REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
15.Feb.00

3. REPORT TYPE AND DATES COVERED
THESIS

4. TITLE AND SUBTITLE
THE EFFECTS OF DISPLAY HIGHLIGHTING AND EVENT HISTORY ON OPERATOR DECISION MAKING IN A NATIONAL MISSILE DEFENSE SYSTEM APPLICATION

5. FUNDING NUMBERS

6. AUTHOR(S)
2D LT SMITH MELISSA A

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITY OF ILLINOIS AT URBANA

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
THE DEPARTMENT OF THE AIR FORCE
AFIT/CIA, BLDG 125
2950 P STREET
WPAFB OH 45433

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
FY 00-63

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Unlimited distribution
In Accordance With AFI 35-205/AFIT Sup 1

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

20000307034

14. SUBJECT TERMS

15. NUMBER OF PAGES
68

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT

18. SECURITY CLASSIFICATION OF THIS PAGE

19. SECURITY CLASSIFICATION OF ABSTRACT

20. LIMITATION OF ABSTRACT

THE EFFECTS OF DISPLAY HIGHLIGHTING AND EVENT HISTORY ON OPERATOR DECISION MAKING IN A NATIONAL MISSILE DEFENSE SYSTEM APPLICATION

BY

MELISSA ANNE SMITH

B.S., United States Air Force Academy, 1998

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 1999

Urbana, Illinois

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

THE GRADUATE COLLEGE

JULY 1999
(date)

WE HEREBY RECOMMEND THAT THE THESIS BY

MELISSA ANNE SMITH

ENTITLED THE EFFECTS OF DISPLAY HIGHLIGHTING AND EVENT HISTORY ON OPERATOR DECISION MAKING IN A NATIONAL MISSILE DEFENSE SYSTEM APPLICATION

BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF MASTER OF SCIENCE

Director of Thesis Research

Head of Department

Committee on Final Examination†

Chairperson

† Required for doctor's degree but not for master's.

ABSTRACT

A proposed display for the National Missile Defense (NMD) task was developed to aid operator decision making. Subjects were required to monitor a simulated battle, consisting of launches of enemy missiles against the U.S., and counter-launches of defensive missiles against these incoming warheads. Defensive missiles were not perfect at destroying targeted enemy missiles, there was an estimated probability of .20 that a defensive missile will miss its assigned target. The risk associated with the probabilistic outcomes was displayed to the operator as a distribution in part of the display. The counter-launches were accomplished by a fully automated system, with the human operator as a monitor. The subject controlled a pool of reserve missiles (which are limited in number), not included in the system automation, which were deployed when the subject determined that the threat called for such action. Subjects had to make risk-resource tradeoffs concerning the risk associated with the threat and the limited resources of the reserve missiles. Twenty military subjects saw 40, two minute scenarios, with two enemy launches of six missiles each. They were required to respond as to how many reserve missiles they wanted to withdraw at four times during each scenario. The independent variable of display consisted of the highlighting of one of the three possible outcomes (best case, worst case, expected case) on the risk-resource display, or no highlighting, for a total of four levels. The trend variable consisted of different outcomes of the success of each counter-launch against the enemy launches. No significant effect of display highlighting was observed, possibly due to experimental considerations. A significant effect of trend was observed, with more reserve missiles withdrawn as time went on during a scenario, as well as the situation became more threatening (more incoming enemy missiles). Further analysis revealed the existence of recency and primacy effects (specifically contrast effects) on the number of missiles withdrawn.

TABLE OF CONTENTS

1. INTRODUCTION

A disturbing, yet realistic threat to the United States is the launch of missiles towards the country, from one of the hundreds of countries that possess the technology to do so. Surprisingly, a comprehensive system for the defense against ballistic missiles has not yet been implemented, even with advances in modern day technology. Since the role of the human operator in the National Missile Defense (NMD) system will be primarily one of monitoring, the development of an effective defense should incorporate theories of human cognition, and, more importantly the results of realistic, applied testing. Unfortunately, literature pertaining to such an applied setting for this system is sparse. This study seeks to fill the gaps in an effort to provide data relevant to the development of the NMD system.

In an assumed scenario, a number of missiles (containing multiple warheads) will be launched at the United States from abroad. The NMD system automatically targets these and allocates defensive missiles (Ground Based Interceptors) to destroy the incoming warheads. The success of the Ground Based Interceptors (GBIs) is not guaranteed; rather it is probabilistic, and only a finite number of GBIs are available. Probabilities of the varying levels of success are computed by an automated system, based on a series of complex algorithms, and only the final results are displayed to the operator. The human operator, given this data, can intervene and choose to withdraw additional, reserve GBIs to be launched at incoming warheads. The role of the human represents the final safeguard in the system and is crucial, since new intelligence information can arise which is not considered by the automation algorithms.

The decision of the human to intervene or not to intervene must be made under severe time pressure, and with the possible consequence of nuclear holocaust in mind. This risk-resource tradeoff is a very important part of the human operator's task and makes relevant several major psychological concerns. If the operator chooses not to withdraw enough reserve GBIs from the pool of reserve GBIs he or she controls, the US may not have an adequate defense for the incoming missiles. On the other hand, if the operator chooses to withdraw too many reserve GBIs, not enough GBIs will remain to combat future threats of enemy warheads launched.

One prototype of a graphical NMD display that is being developed provides the operator with several pieces of information that he or she must interpret in order to make GBI withdrawal decisions. The display consists of a representation of the incoming warheads as a function of time, as well as a Risk-Display, which will be the topic of investigation for this thesis. Since each GBI is estimated to be about 80% successful in hitting the warhead it is targeted to hit, additional risk exists because not all GBIs may hit their assigned target warheads. This risk is shown in the form of a graphical distribution based on the binomal distribution, called the Risk-Display. This display shows a probabilistic distribution portraying the number of GBIs needed to handle the best case situation (where all warheads are hit by all GBIs fired), the worst case situation (where all warheads get by or "leak" through the GBI defenses) and the expected case situation (a value characterizing the central tendency of the distribution).

An example of the Risk-Display is shown in Figure 1 below. The scale on the y-axis is in terms of the number of GBIs required to handle the current threat. Within the curve, the x-axis represents the actual probabilities associated with each outcome.

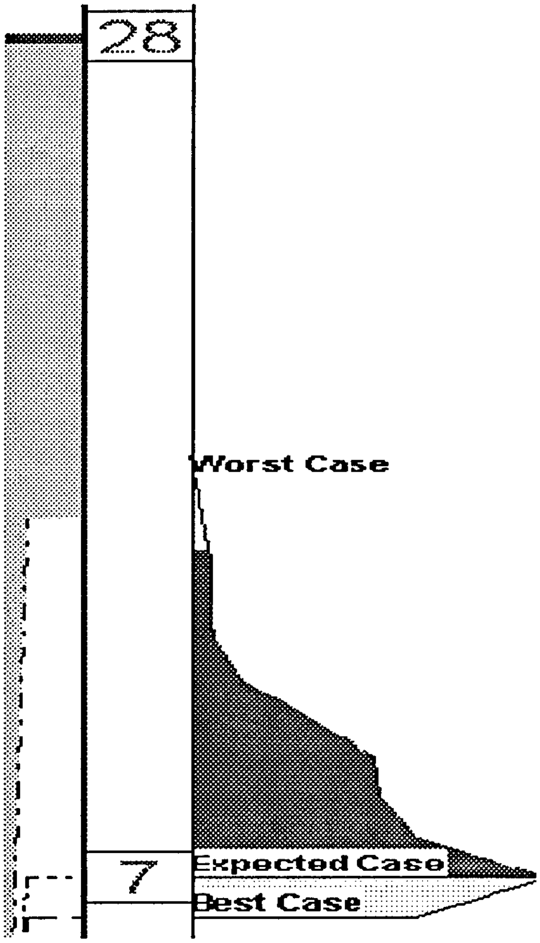


Figure 1. Sample Risk-Display

The three probabilities, and thus the shape of the distribution, will change with the number of enemy warheads still en route towards the United States at any given time. While the shape will also change with the assumed probability of success of each GBI, in the current case this is held constant at p=.80.

Preliminary results from evaluating operator monitoring behavior of the NMD system in simulations indicate that people tend to intervene in the system automation more than they should (M. Barnes, personal communication, 1999). This "trigger happy" tendency creates a very serious problem for the NMD task. Here, the consequences for inappropriate human intervention are very severe. If the operator does not trust the automation and withdraws additional GBIs, the supply of GBIs will be unnecessarily decreased, leaving less for the defense against future enemy launches.

The tendency of humans to over-intervene is a human performance issue that must be investigated thoroughly before a satisfactory NMD system can be designed. The "trigger happy" tendency in the NMD task will be considered in the contexts of human performance with automation and decision making with probabilistic information. Flaws in monitoring performance are not uncommon among humans when dealing with automation. As previously mentioned, the human in the NMD system takes on the role of the system monitor, and it has been shown that humans are not generally very effective monitors (Parasuraman, 1986). The fact that the majority of the task requires monitoring the behavior of an automated system may degrade operator decision making due to a lack of understanding of the complex algorithms performed by the automation.

Research has also shown that humans sometimes have a difficult time interpreting and applying probabilistic information when making decisions, which may help explain the tendency to intervene (Edwards, 1968; Schipper and Doherty, 1983; Schum, 1975). For instance, while the worst case situation is a very low probability event that requires the launching of many GBIs (and therefore a strong need to withdraw reserves), the probability of this event may be greatly overestimated when making decisions, an overestimation resulting from the salience in memory associated with rare events, an issue to be discussed later. This salience would result in the desire to withdraw more GBIs than necessary.

In probabilistic environments, the most accurate way to make decisions is by evaluating the probabilities associated with the potential outcomes. Research indicates that people do not typically make decisions this way in everyday life (Edwards, 1968). Not enough weight is given to the actual probabilities associated with information. For instance, Schum (1975) found that when jurors were presented with an eyewitness known to be unreliable, they failed to account for this unreliability when making their decisions about a verdict, and they treated all witnesses as equally reliable.

Within the research on probabilistic decision making, many biases have been shown in the literature depending on the type, amount and order of information presented. For instance, the results of a GBI launch (number of hits or misses of GBIs against enemy warheads) may influence subsequent decision making, even though the probabilities associated with each possibility are not necessarily altered by the results of prior launches.

To understand these problems in an attempt to alleviate them, we must explore the relevant literature in each of these areas. First, a review of the relevant literature concerning human interaction with automation will be presented. This is followed by a discussion of biases in probabilistic decision making, including errors in the interpretation of probabilities and influences the display may have in inducing such biases. Finally, a summary for the literature review, and a general introduction to the experiment will be presented.

1.1 Human Interaction With Automation

While meant to reduce the workload and/or improve performance of human operators, the introduction of automation can actually result in less than optimal performance from the human (Parasuraman and Riley, 1997). When automation is added to a task previously performed by humans, the role of the human changes from that of the primary operator to that of a system monitor, and as a result, also changes the cognitive demands of the task. Despite the fact that the automation was intended to ease the cognitive demands of the operator, these new demands may sometimes be greater than the original demands without automation.

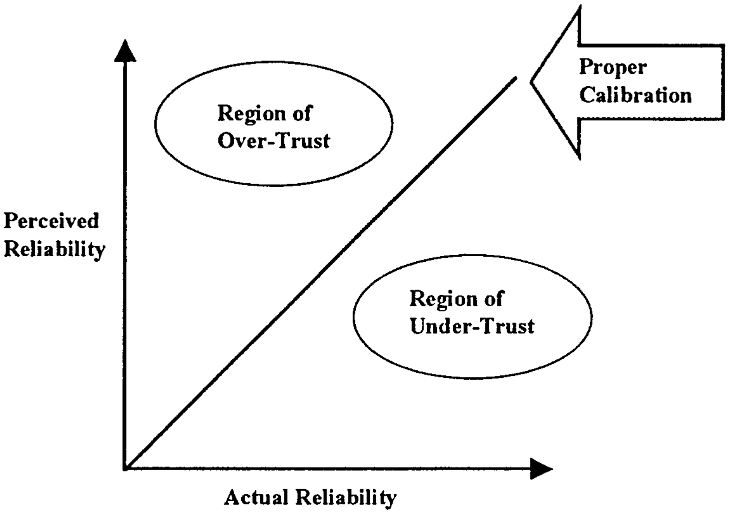


Figure 2. Human Trust Calibration

Figure 2 (Gempler and Wickens, 1998) summarizes the calibration function relating human trust with automated systems reliability. As shown in the figure, when an automated system is *perceived* to be completely reliable by the human operator (shown along the y-axis), yet is not completely reliable (shown along the x-axis), the human operator can be described as over-trusting the automation. This is the region shown in the top left oval of Figure 2. Over-trust of automation occurs when the operator assumes that the automation will function correctly, and consequently fails to monitor it with sufficient vigilance and to intervene when problems arise (Parasuraman and Riley, 1997). This over-trust increases as the actual reliability of automation increases (McFadden, et al, 1998) and as the human is progressively removed from the system, i.e. relegated to the role of a passive monitor (Parasuraman, Mouloua, Molloy, and Hilburn, 1996; Wickens and Kessel, 1979).

On the other hand, when an automated system is perceived to be less reliable than it actually is, the human operator tends to under-trust the automation (bottom right oval of Figure 2). Such behavior can result in the automation not being used to its full potential and possibly its replacement by less than perfect human performance i.e., the "trigger happy" phenomenon reported anecdotally in the NMD task. The literature has shown that a person's under-trust of automation can result from any of three factors: poor understanding of automation processes, poor automation performance and a user's overconfidence in his or her own abilities. First, inadequate understanding of the automation processes has been found to be a factor in under-trust in several studies. For example, Sarter and Woods (1995) observed this inadequacy when studying commercial airline pilots flying with automated flight management systems. Second, under-trust is also related to past poor performance by the automation, such as frequent false alarms, as shown by Muir and Moray (1996) and Parasuraman and Riley (1997).

Third and finally, overconfidence in human abilities has been shown to be a factor in under-trust. If people assume that their own performance will be better than the automation, they will fail to utilize automation, even in cases where it performs better than the person (Riley, 1996). As an example, Liu, Fuld and Wickens (1993) had subjects complete a task involving the assignment of customers to check-out lines in a supermarket. This assignment would either be done manually, or through an automated system, where the subject was in charge of monitoring the system. Unbeknownst to the subject, the monitoring situation was simply a recording of his or her own performance in the manual situation. The results indicated that people trusted their own performance over that of the automation (i.e. were less likely to detect errors), even when the performance in the two conditions was exactly the same. Under-trust in the NMD system may be a product of the general overconfidence of humans in their own cognitive abilities, or of perceived unreliability of the system.

100% reliability in the NMD system is not possible given that the Ground Based Interceptors are estimated to have only an 80% probability of hitting their assigned warhead targets and furthermore that the automation algorithms may not always effectively account for this GBI inaccuracy. As mentioned previously, initial data from prototype evaluations of the NMD system (M. Barnes, personal communication, 1999) revealed that the human operators tend to behave in the region of under-trust. In order to make human operator performance more optimal, we must determine how to calibrate operator trust of the NMD automation. Most research on human interaction with automation to date has focused on the region of over-trust. The present study seeks to examine the region of under-trust in more detail, and to uncover the mechanisms of human decision making that create mistrust.

The diagonal line in Figure 2 indicates ideal calibration between the actual reliability of the automation and the human operator's perception of the automation reliability. Merlo, Wickens and Yeh (1999) suggest that a display that depicts the reliability of automation may improve trust calibration. The proposed NMD display attempts to take into account the less than perfect reliability of the GBIs. This reliability is displayed to the human operator in the form of a graphical distribution of probabilistic outcomes, an example of which was shown in Figure 1. However, displaying such probabilistic information poses new problems since humans are not always adept at interpreting this type of information (Wickens, 1992). We now turn to examining biases in probabilistic decision making which will be discussed to further understand the role of the human operator in the National Missile Defense system.

1.2 Biases in Probabilistic Decision Making

The National Missile Defense Operator (NMDO), is required to process several pieces of very complex information that are presented over time, and make a decision on whether to intervene or not intervene in the system automation by withdrawing or withholding reserve GBIs. Wickens, Gordon and Liu (1998) identified three stages of decision making from an information-processing perspective: attention, diagnosis and choice. These aspects of the NMDO's decision making task are represented in Figure 3:

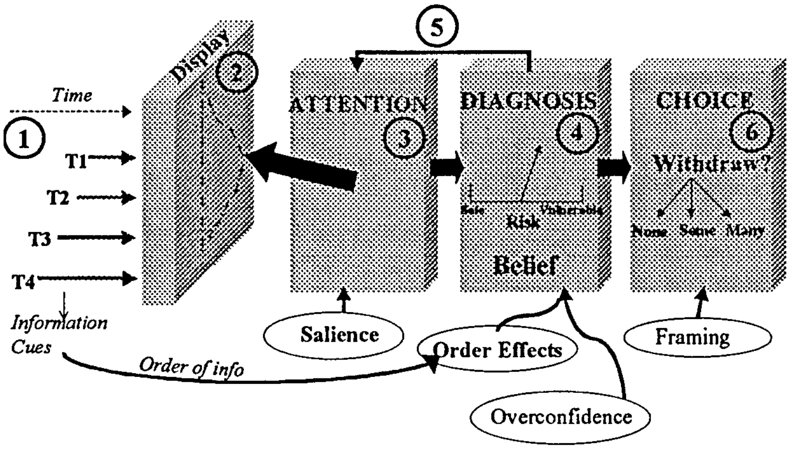


Figure 3. NMD Decision Making Framework

Figure 3 depicts the stages in decision making from left to right in the NMD task, as gray blocks. The sequence of processes is shown with arrows in the Figure, with each event labeled numerically in circles. Biases or heuristics that are expected to come into play at each stage are shown in ovals. Particularly relevant biases, which will be explicitly examined during the present study, are shown with bold text within the ovals. At far left in the Figure, depicted by the arrows, are shown multiple sources of information about incoming enemy warheads, which arrive to the operator over time (T1-T4). This information (the number of incoming warheads and the number of destroyed warheads) has predictable effects on the vulnerability or risk of the United States, and these effects are shown to the operator in the form of a probabilistic Risk-Display (2) of outcomes (best case, worst case and expected case probabilities, see also Figure 1). At process 3, the operator allocates his or her attention to different features of the probabilistic display and then in process 4, forms a subjective belief (diagnosis) of the vulnerability or safety of the United States. This belief may then drive operator attention back to the display to seek additional information (5) and will certainly lead to a choice (6) of whether or not to withdraw additional missiles from a reserve pool (i.e. whether or not to "over-ride the automation"). If the choice is to withdraw, a decision is made of how many to withdraw based on the assessed vulnerability. Within this framework, several decision making biases and heuristics have been identified by researchers (as shown by the ovals), many of which may affect the NMDO's cognitive processing, and two of which, salience and order effects, will be explicitly examined in the present experiment.

1.2.1 Graphic Probabilistic Displays

We made a decision at the outset of this study to develop and provide a graphic probabilistic display to help the NMDO. This decision is consistent with three findings in the literature: 1) providing predictive information for tasks where prediction is involved tends to aid decision making, 2) graphic displays can improve performance and 3) graphical displays of probability tend to aid decision making over numeric displays of probability. These findings will be discussed in the following section.

First, in the NMD task, planning/forecasting is involved since the number of GBIs to withdraw must be anticipated by the human operator. This planning is based on predictive information as provided by the NMD Risk-Display. When people are asked to predict future trends based on present and past values, they generally do not perform well, and tend to give predictions that are overly conservative (Wickens, 1992; Waganaar and Sagaria, 1975). In order to alleviate this deficiency in human decision making, when people are given information as a preview of likely future events, they tend to behave more optimally (Wickens, 1992). One method of providing information as a preview of likely future events is accomplished by predictive displays. Wickens and Morphew (1997) tested types of predictive displays in decision making for an aviation task. The predictive displays increased pilot performance and decreased pilot workload over the non-predictive displays. Furthermore, Gempler and Wickens (1998) evaluated an uncertainty representation in the predictive display and found this to be feasible; while it did not improve performance, it did reduce workload. In the NMD task, showing the probabilities of three anticipated outcomes (worst, best and expected as shown in Figure 1) is analogous to presenting a predictive display with an explicit representation of forecast uncertainty since these probabilities show the likelihood of outcomes to follow. Since predictive displays seem to guide decision making, for the NMD task we might expect operators to base their decisions to withdraw GBIs on the particular probability (worst case, best case or expected case) that is highlighted.

Second, graphically displaying the actual probabilities that human operators must take into account during decision making is one method of helping humans with interpreting probabilistic information. The value of graphical displays has been evaluated in several studies. Pitz (1980) suggests that graphic displays of probability (and other quantitative information) may result in better decision making since such displays are hypothesized to be interpreted by fast perceptual processes, rather than by the slower processes involved in encoding the symbols inherent in numeric representations. Graphical displays of uncertainty in a prediction task were also shown to be helpful by MacGregor and Slovic (1986).

Third, the literature has shown that graphical displays of probability information help decision making over the numerical display that is characteristic of that currently used in the NMD prototypes. Andre and Cutler (1998) showed that displaying uncertainty or probability could help performance over no display of uncertainty and that graphically displaying uncertainty lead to the better performance than displaying uncertainty numerically. Kirschenbaum and Arruda (1994) evaluated different methods of displaying probability information for the spatial task of predicting the location of submarines. Their results indicated that graphic displays of probability *did* result in better predictions than verbal representations.

Schwartz and Howell (1985) used a simulated hurricane-tracking scenario to evaluate decision making. They varied the display of position history (numeric or graphical) of a hurricane and required subjects to make a decision (whether or not to evacuate a populated area) at various points throughout the scenario. They also added a decision aid in some conditions by displaying updated probabilities of the likelihood that the storm would hit a populated area at each decision point in the scenario. These probabilities were presented numerically. Their results showed that subjects in the numeric display condition were less efficient (took more time, no increase in accuracy) than the graphical condition. The probability decision aid raised the accuracy of all decisions significantly. When the same experiment was conducted using display format as a within subjects variable, graphical displays of time history resulted in significantly greater performance than numerical displays.

Stone, Yates and Parker (1997) examined the effects of various ways of displaying low-probability risk information on risk-taking behavior. They presented subjects with information on a brand of tires said to reduce risk of blowout over the standard brand and information on a brand of toothpaste said to reduce the risk of gum disease over a standard brand of toothpaste.

Both risk-reducing brands had higher prices than the standard brands. The risk information was presented to subjects was presented in numeric format of expected injuries, or in different graphical formats. Their results indicated that graphical representations decreased risk-taking behavior over numerical representations of risk. This means that subjects were more likely to purchase the more expensive, yet safer, products even though their probability of occurrence was very low.

Thus, the literature has provided evidence for our display selection. First, providing predictive information can help performance for tasks where prediction is required (Wickens, 1992; Wickens and Morphew, 1997; Gempler and Wickens, 1998). Second, the value of graphical displays for improving performance has been shown in several studies (Pitz, 1980; MacGregor and Slovic, 1986). Third, graphically displaying probability or uncertainty has been shown to aid decisions over numeric or verbal representations (Stone, Yates & Parker, 1994; Schwartz and Howell, 1985; Kirschenbaum and Arruda, 1994; Andre and Cutler, 1998).

For the NMD task, risk is displayed in both the graphical form of the probability distribution, and the numeric form of the expected number of GBIs required to handle the threat, as shown in Figure 1. Highlighting the numeric form may reduce the subjects tendency to focus on the low probability worst case scenario and withdraw more GBIs than required. The issue of what is and is not highlighted or noticed in a complex decision display such as that in Figure 1, makes relevant the issue of *salience* and its effect on decision making.

1.2.2 Salience

As discussed earlier, a display such as the Risk-Display shown in Figure 1 can provide aiding to probabilistic decision making. However, such a complex, multi-element graphic display, particularly when used under time pressure, will challenge selective attention, possibly causing users to overweight or over-attend to certain parts at the expense of others. Research indicates that highly salient information tends to influence decisions more than it should in general (Payne, 1980) but even more so under conditions of time pressure (Wickens and Hollands, 1999). In the NMD task, since the information presented is complex and time pressure is involved, operators may tend to make decisions based on the most salient pieces of information without accounting for the other relevant pieces of information that are presented to them. Past research has revealed that attention to multi-element decision displays is driven heavily by salience, but salience itself can be defined in two contexts: salience of the mental representation for a signaled element and physical salience of a display item. We discuss each of these in turn:

The salience in memory associated with rare events with extreme consequences, appears to result in an overestimation of the probability of their occurrences. In the NMD task, the display provides probabilistic information to the operator for three outcomes (worst case, best case and expected case, see Figure 1). The worst case probability is defined as the likelihood of the event where all enemy warheads fired at the United States survive, or "leak," through the GBI defenses. This worst case probability is a very low probability, rare event. For example, if six enemy warheads are launched against the United States, then given the estimated .80 value of destroying each one, the calculated probability of all six of them surviving the initial defensive launches and hitting their targets is only *.000064*, but a number of such a small magnitude is difficult for most humans to comprehend accurately. For example, it has been shown that people tend to overestimate the occurrence of very rare events (Tversky and Kahneman, 1981).

One example of the overestimation of low probabilities has been shown with subjects estimating statistical values of a data set. Numerical estimates of variability seem to be influenced by highly salient, or extreme (and rare), members of the data set, instead of being fairly based on all values of the data set (Pitz, 1980). The overestimation of the probability of rare events also leads to behavior reflected by the purchasing of insurance against accidents or disasters, or high stakes gambling. In both cases, the events in question (accident or disaster and winning large amounts of money) are quite rare, yet, because of their salience humans tend to overestimate their probabilities of occurrence and purchase insurance policies or place bets. This happens because the subjective expected gain (saving property or winning money) or loss (accident or losing money bet) is higher than the objective expected gain or loss, as is evidenced by the financially successful insurance and gambling industries throughout the world. In the NMD task, the low probabilities are associated with the worst case situations. These situations, although extremely rare, may be quite salient to the operator as is the thought of a costly car accident or natural disaster to the potential insurance customer, or the thought of winning a million dollars is to the potential gambler. This salience would help to explain the tendency of operators to over-intervene in the automation, as shown by preliminary simulation results. Salient risks are also overestimated by humans, despite their low probability. For instance, Combs and Slovic (1979) showed that people over-estimated the risk of highly publicized events in the media such as terrorist bombings or plane crashes. Additionally, Slovic (1987) found that people estimated the risk of dying in a plane crash (rare, yet salient event) as higher than the risk of dying from a fall in the home, even though these risks are actually reversed. In summary, the very low magnitude of the probabilities associated with rare events is generally not taken into consideration when people make their decisions.

The overestimation of probabilities of rare events may be best summarized by a hypothetical function created by Tversky and Kahneman (1981) to describe human decision behavior, shown in Figure 4 below:

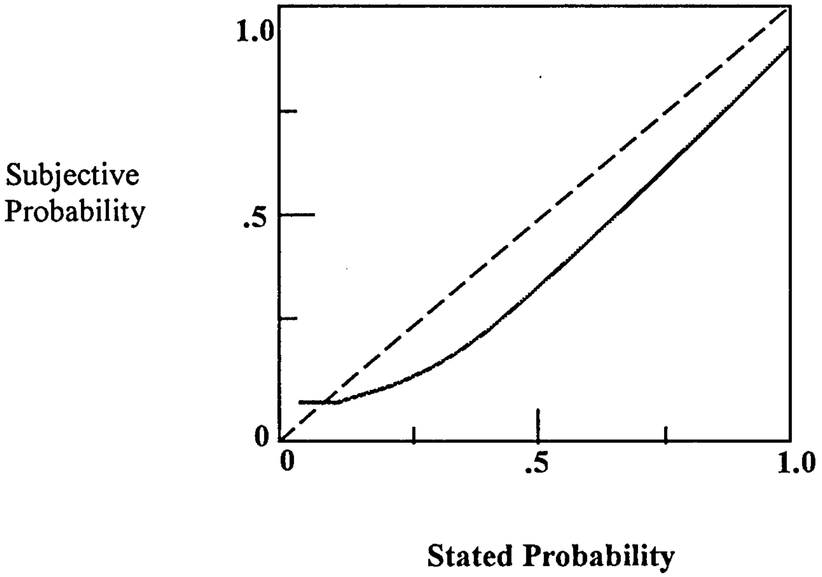


Figure 4. Weighting function for human probability assessment (Tversky and Kahneman, 1981).

As shown in the figure, human interpretation of probabilities (as shown by the red diagonal line) is exaggerated for rare events (with small probabilities) and is also somewhat insensitive to changes in probability values at low probabilities, as shown by the flattened out portion of the curve for the lower probability values. The above literature provides a partial explanation for the operator's tendency to intervene in the automation as shown in the preliminary results of the NMD simulations (M. Barnes, personal communication, 1999). The salience of the worst case probability would cause the operators to plan for that outcome by withdrawing more GBIs from the reserve, even though the worst case is a very low probability event.

Given this knowledge about the human tendency to overweight in memory the likelihood of low probability but salient events, we can now focus our efforts on trying to remediate the tendencies that would lead to poor performance in the NMD task (the desire to withdraw more GBIs than necessary). To do this, we consider the second definition of salience for the present study: the perceptual salience of display properties. How complex information is displayed can greatly affect operator decision making (Wickens and Hollands, 1999). The brightest, most colorful, or visibly located display properties are typically the most salient, leading human operators to process the content contained in these highly salient display parameters over that of less salient display parameters (Wickens, 1992; Payne, 1980). Wallsten and Barton (1982) examined one aspect of perceptual salience by manipulating the position of information cues within a display. They found that the cues presented at the top of the display (more salient) were selectively processed even though they were contained equal or less diagnostic value than the cues at other positions within the display. This selective processing of perceptually salient information can bias decision making (Wickens and Hollands, 1999).

Kaplan and Simon (1990) found that decision making is improved if the critical attributes of a decision are perceptually salient. Montgomery (1999) found similar results in a study on human sensitivity to variability of information. In this study, subjects had to interpret a graphic display which portrayed information about the variability of several sources of data. They then had to decide which source had higher reliability (lowest variability). Subjects who saw a display with the most diagnostic information highlighted performed better than those who did not. Since we would like people to behave more optimally with the NMD system, or make their decisions based on expected case probabilities rather than the memorably salient worst case, highlighting the expected case portion of the display should lead to more appropriate decisions, whereas highlighting the worst case probability should lead to the tendency to withdraw more GBIs than are required.

Schwartz and Howell (1985) summarize the possible effects of display manipulations on human decision making as follows:

> " . . . display format can affect decision performance in very subtle ways. Not only can it bring out aspects of a data set (such as trend information) that are otherwise difficult to perceive—the well-established and relatively obvious 'compatibility' phenomenon—it can alter the decision maker's whole approach to information processing. In a sense it can alter his/her processing 'set.'"

A review of the literature revealed no decision making study that has systematically examined the processing of best case, worst case and expected case probabilistic data within a distribution such as that shown in Figure 1, nor how this processing (and the distribution of attention) might be altered by changing the nature of the display representation. In the present study, participants are presented not only with the worst case probability information, but also the best case and the expected case probabilities. We seek to determine how highlighting one of these probabilities in a display could influence operator decision making.

1.2.3 Biases in the Diagnosis Stage of Decision Making: Overconfidence

In the diagnosis stage, where operators formulate their beliefs about the current situation, the operators' confidence in their own decision-making ability may affect their decision. For instance, if an operator feels that he or she can make GBI allocations better than the automation, he or she will be likely to intervene. Several studies have shown that people are generally overconfident in their state of knowledge (see Wickens and Hollands, 1999 for a review) and people tend to report high confidence levels in their assessments or predictions, regardless of the correctness of their answers (Kleinmuntz, 1990). This overconfidence has also been shown in judgments of abilities, such as eyewitness memory and knowledge of facts (Wells, Lindsay and Ferguson, 1979; Fischhoff, Slovic and Lichtenstein, 1977; Bornstein and Zickafoose, 1999), as well as in human interaction with automation as discussed earlier (Kleinmuntz, 1990).

Overconfidence is also present when humans must integrate information over time, such as in the NMD task. The critical issue here is how confidence evolves over time, i.e., if people are overconfident in their diagnosis, they may stop their search for information early and make a premature decision based on incomplete information (Wickens and Hollands, 1999). In the NMD task, if an operator is overconfident, they may both be more likely to intervene and to stop their information seeking early, and make decisions based on initial information presented, similar to the primacy effects to be discussed in the following section.

1.2.4 Biases in the Diagnosis Stage of Decision Making: Primacy and Recency

One piece of information available to the NMD officer is the results of previous GBI launches against incoming warheads from earlier engagements. The operator can see on the display how successful the automation's GBIs were in destroying enemy warheads from earlier enemy launches within the battle. Studies on biases in probabilistic decision making has shown that that past events can shape present decision making when information must be integrated over time (Adelman and Bresnick, 1992; Tversky and Kahneman, 1974; Hogarth and Einhorn, 1992). However, in the case where new information entirely updates old information, then past events (old events) should have no relevance, and recency is an appropriate heuristic for making decisions. In the same case, if the old information is considered more relevant than the new information, a primacy effect exists. The primacy effect can take two main forms: anchoring effects or contrast effects. Anchoring occurs when the later situation is evaluated the same as an earlier situation, regardless of evidence to the contrary. What is unclear is how these heuristics and biases will play out in terms of the missile defense task.

In the NMD task, the operator must perceive information and integrate it over time in order to make accurate decisions. However, integrating information over time is difficult and consequently, the human operator often takes advantage of systematic heuristics. One of these heuristics is called anchoring, describing a type of primacy effect. Here, diagnoses are made while using the initial piece of evidence as an anchor (Tversky and Kahneman, 1974). That is, the first piece of information presented or discovered has a stronger effect on the final diagnosis than do subsequent pieces of information. Literature on primacy effects in decision making for an applied setting is relatively scarce. One relevant applied study by Tolcott, Marvin and Bresoick (1989) examined the decision making of Army intelligence analysts. These analysts were given multiple pieces of information concerning the location of an enemy force during a simulated battle. The analysts developed an initial hypothesis with this information and subsequently were found to give significantly higher weights to evidence that was consistent with these initial hypotheses. The results of this study suggest that anchoring effects may be observed during applied, dynamic decision making tasks, such as those of a National Missile Defense human operator.

Anchoring is one example of primacy, in which later diagnoses are consistent with earlier information. For example, if an earlier evaluation was of a poor situation, then later evaluations will also be biased toward the negative, independent of subsequent evidence. However, primacy may also be manifest as a "contrast effect." This effect is defined as the opposite of the anchoring effect. For example, if an earlier evaluation was of a poor situation, then later evaluations will be biased positively, independent of the evidence. This form of primacy reflects a trend effect such that evaluations depend on the rate of change between past and present evaluations.

While anchoring involves decisions made by overly weighting an initial piece of information, the recency effect in information integration and diagnosis describes a heavier weighting of the most recently observed piece of information. The order effects of primacy and recency can also appear together in some combination resulting in middle pieces of information being weighted less than the other pieces of information in decision making (Hogarth and Einhorn, 1992). Hogarth and Einhorn (1992) examined circumstances under which recency versus anchoring would occur. They found that for simple tasks, where a judgment is made only once, after receiving all information, anchoring is observed most often. For more complex tasks, where overt judgments are required several times during the sequence of presentation of information relevant for a diagnosis, and the information is therefore processed in steps, the recency effect is observed most often. The NMD risk assessment task is definitely a complex task, so we might expect recency to be a factor in operator decision making. Furthermore, the operator must make repeated assessments after each new piece of arriving information, also inviting recency. The particular NMD scenario is one in which each new piece of information is assumed to update and replace prior information. Hence, it is a scenario in which recency *is* optimal, and anchoring is clearly a non-optimal bias. Our interest in this study was to determine the extent to which anchoring had any effect on performance. That is, the extent to which an assessment at time T was influenced by the nature of evidence presented at an earlier time.

1.2.5 Biases in the Choice Stage of Decision Making: Framing

One limitation of human decision making is framing. In general terms, framing describes how the way in which a choice is presented can influence the decisions made. For example, in choices involving gains people tend to be risk averse, whereas in choices involving losses, people tend to be risk seeking, even though the probabilistic outcome in both choices may be equivalent (Tversky and Kahneman, 1981; Carroll, 1980; Puto, Patton and King, 1985; Schurr, 1987; McNiel, Pauuker, Sox and Tversky, 1982).

In the NMD system, "risky" operator behavior is defined as intervening in the automation, and withdrawing more GBIs from the reserve. From the operator's perspective, this behavior would be deemed "conservative" since more GBIs are withdrawn to reduce leaking warheads and protect the United States. From the operational perspective however, any instance where the human intervenes to override the automation is considered to be "risky" behavior. Framing has also been shown in decision making for dynamic systems. Nygren (1997) showed that the framing of instructions (losing or gaining performance points) in a multitask dynamic system influenced decision making strategy such that subjects in the negative framing group performed differently than those in the positive framing group. In the NMD task, the way in which the instructions for the task are presented to the operator for example in terms of saving or losing human lives, or saving or losing GBIs for future use, may cause framing effects. This bias is guarded against in the present study by providing all subjects with the same task instructions that contain both positively and negatively framed wording (further discussed in methods section).

1.3 Literature Summary and Task Overview

The literature on probabilistic decision making and human interaction with automation identifies several biases that may be relevant to the decision stages of the NMD task. Preliminary investigations have shown that operators in the NMD task have a tendency to intervene too much in the automation (M. Barnes, personal communication, 1999). A review of literature on human interaction with automation (Riley, 1996; Liu, Fuld and Wickens, 1993) reveals that under-trust of the automation, due to operator overconfidence and/or lack of understanding of automation processes, may be responsible for the tendency to withdraw more GBIs than are optimal. During the decision making process, salience of display parameters may bias decision making (Payne, 1980). Also, the order in which information is presented may change decision making such that information presented last in a sequence (recency), during a multi-step task, is weighted more than the other pieces of information (Einhorn and Hogarth, 1992); but other circumstances may lead to primacy instead.

We now turn to an overview of the experimental NMD simulation that we will employ to provide context in which we will examine the nature of order effects and the influence of display salience. In the simulation, information is presented to the human operator that must be processed in limited time with severe consequences. The operator views a display which shows enemy warheads as icons moving across the screen as they travel across the earth towards the United States. As soon as the automation identifies these enemy warheads and determines their intended targets, it launches our GBIs at each of the incoming warheads. This allocation of GBIs to warheads is also shown on the operator display screen. The results of our counter-attack of GBIs for each launch of enemy warheads, are also shown to the operator. The total number of GBIs available in the immediate pool is also presented, as well as a display that shows the expected case, worst case, and best case value (all in terms of number of GBIs necessary for defense in the relevant case) as they change with the course of events, as shown in Figure 1. The operator monitors the automated system, but also has control over a pool of 15 “reserve GBIs,” not included in the calculations of the automation. We chose to incorporate 15 reserve GBIs to be consistent with the number used in the actual version of the software. At four specific decision points at which new information becomes available, the operators can choose to withdraw and deploy these reserve GBIs when they determine that the risk warheads hitting the

United States has reached what they feel to be an unacceptable level. However, they are cautioned as to the value of these reserves for others, and that they should not be squandered needlessly. In the actual NMD system, non-optimal operator behavior would be very costly. If operators chose to withdraw too many reserve GBIs, our defense against a future threat would be severely compromised since these reserve GBIs are limited in number. If not enough were withdrawn, our defense against the current threat could be compromised. A possible subject behavior might have been to use all GBIs each scenario, since 15 were given in each scenario. The importance of conserving GBIs, based on possible future threats in the actual system, was emphasized to subjects in the experiment instructions (Appendix A) as well as during the practice scenario.

In the experiment, the subjects saw 40, two-minute scenarios that contained two enemy launches each. We manipulated the portion of the display in Figure 1 that was highlighted (either the worst case, expected case, best case probabilities, or none) to determine salience effects; and across scenarios we vary the time sequence of success of the launches, in order to determine the existence of order effects. Our display manipulations may help to calibrate the human operators trust in the automation based on the findings of Gempler and Wickens (1998). In the Attention stage of decision making, we expect to see the perceptual salience bias where salient aspects of the display itself get more attention during decision making (Wickens and Hollands, 1999). Thus, highlighting the worst case portion of the display should induce subjects to be more likely to intervene. Highlighting the aspects of the display that would lead to more appropriate decision making could mediate the operator's tendency to intervene. In the Diagnosis stage, we expect to see the operator overconfidence in his or her own decision making. Also in this stage, we predict effects due to the order that the operator receives information over time (Tversky and Kahneman, 1974; Hogarth and Einhorn, 1992). Specifically, we expect to see recency effects where the most recent information influences operator decision making, but we look for any non-optimal contributions of primacy in the form of anchoring effects.

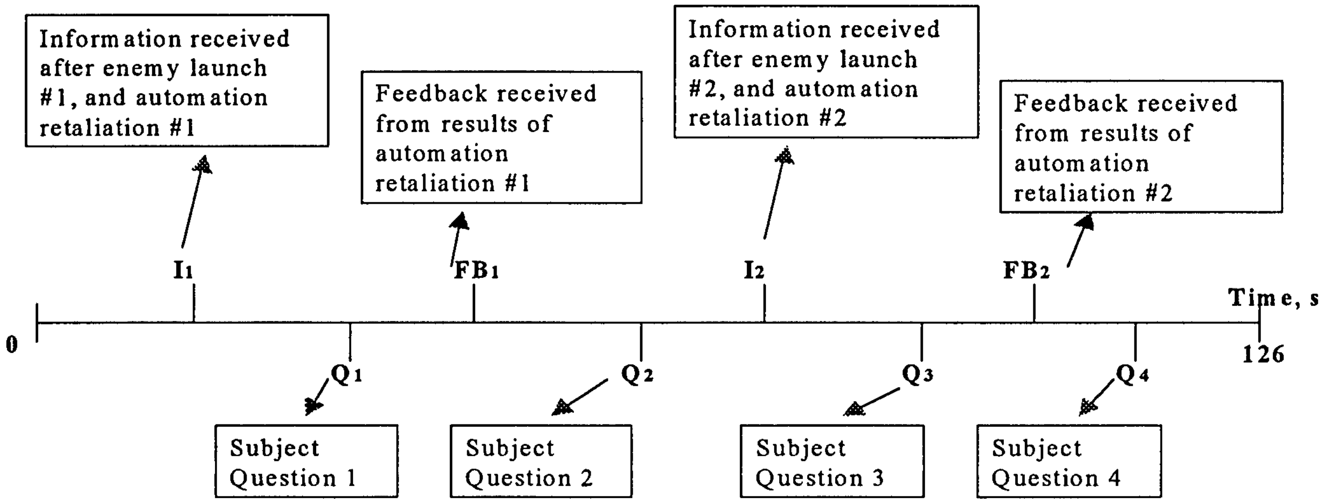
2. METHOD

2.1 Participants

Twenty military personnel participated in the experiment, 17 males and three females. All subjects were United States Air Force and United States Army personnel, four Officers, and 16 Officer Candidates. It should be emphasized that because the NMD system is not actually built there is no actual user population from which to draw subjects. Subjects were paid six dollars an hour for their participation, and were recruited to participate on a volunteer basis. Active duty military personnel declined payment due to administrative restrictions.

2.2 Apparatus and Procedure

The experiment took place on a Windows NT workstation computer, with a 19-inch full color, high-resolution monitor. Subjects sat directly in front of the computer screen, and inputted their answers to questions with the numeric keypad on a normal keyboard. The NMD Visualization software used was a modified version of software developed by Jamieson Christian, TRW, for the Army Research Laboratory. Subjects were introduced to the paradigm through a lengthy series of instructional slides on Microsoft PowerPoint (Appendix A), followed by a two-minute practice scenario with questions allowed. The instructions explained the display components, and showed how the probability distribution was generated. To guard against possible framing effects as seen in Nygren (1997), all subjects were presented instructions which contained both positive (saving human lives) and negative (losing GBIs) framing. Subjects then went through a series of forty scenarios. Each scenario consisted of two enemy launches of six warheads each, and four decision points. Each scenario was 126 seconds long, and stopped automatically. Once one scenario ended, another would show up on the computer monitor automatically, but the subject had to start it him or herself by pressing a play button with a mouse click. The decision points always followed information updates directly. The following figure illustrates the sequence of events for each scenario:



Information received after enemy launch #1, and automation retaliation #1
Feedback received from results of automation retaliation #1
Information received after enemy launch #2, and automation retaliation #2
Feedback received from results of automation retaliation #2
I1
FB1
I2
FB2
Time, s
0
126
Q1
Q2
Q3
Q4
Subject Question 1
Subject Question 2
Subject Question 3
Subject Question 4

Figure 5. Scenario Timeline of Events

Each question asked the subject how many GBIs, if any, he or she wanted to take out of the reserve pool of GBIs to handle the current threat as they interpreted it. Questions remained on the screen for 10 seconds, and no subject failed to answer a question within the time limit. Subjects entered their responses manually on the keyboard. All subjects were allocated 15 reserve GBIs per scenario, and the program kept track of how many they had used for the previous questions. If a subject ran out of GBIs, the only response allowed by the program was "0." In this case, subjects were also asked to report verbally, how many GBIs they would have chose to withdrawn had any been remaining. This was recorded by the experimenter, and used for later analyses to determine the actual pattern of subject responses, in addition to the artificially constrained pattern, with a 15 GBI total for each scenario. All scenarios started with six incoming warheads at I1, could vary in the number destroyed at FB1, depicted an added number of warheads at I2, and could vary in the number of warheads destroyed at FB2. Thus across scenarios the final number of missiles in the air, after FB2 could vary between zero and ten.

The forty scenarios differed in the trend of the two launches (good-good, good-bad, expected-expected, bad-good, bad-bad) as well as the display highlighting (expected case highlighted, best case highlighted, worst case highlighted, no highlighting). After each scenario, the subject was required to verbally report a confidence rating to the experimenter. Subjects took between two and three hours to complete the experiment, and were allowed to take a short break when they were halfway done with all scenarios, as necessary. The completion time differences between subjects were due to individual differences in the lengths of time needed to go through the instructions and practice scenario enough to fully understand the experimental task, as well as desire to take a break or not take a break at the halfway point.

2.3 NMD Displays

Subjects were presented with the full NMD display suite, but instructed to attend only to two parts: the Risk-Display and the Timeline Situation Display, with the Risk-Display being their main source of information for decision making. Other parts of the display suite were visible, but not relevant to the task. Figure 6 shows a sample of the entire NMD display suite which was presented to the subjects during the experiment:

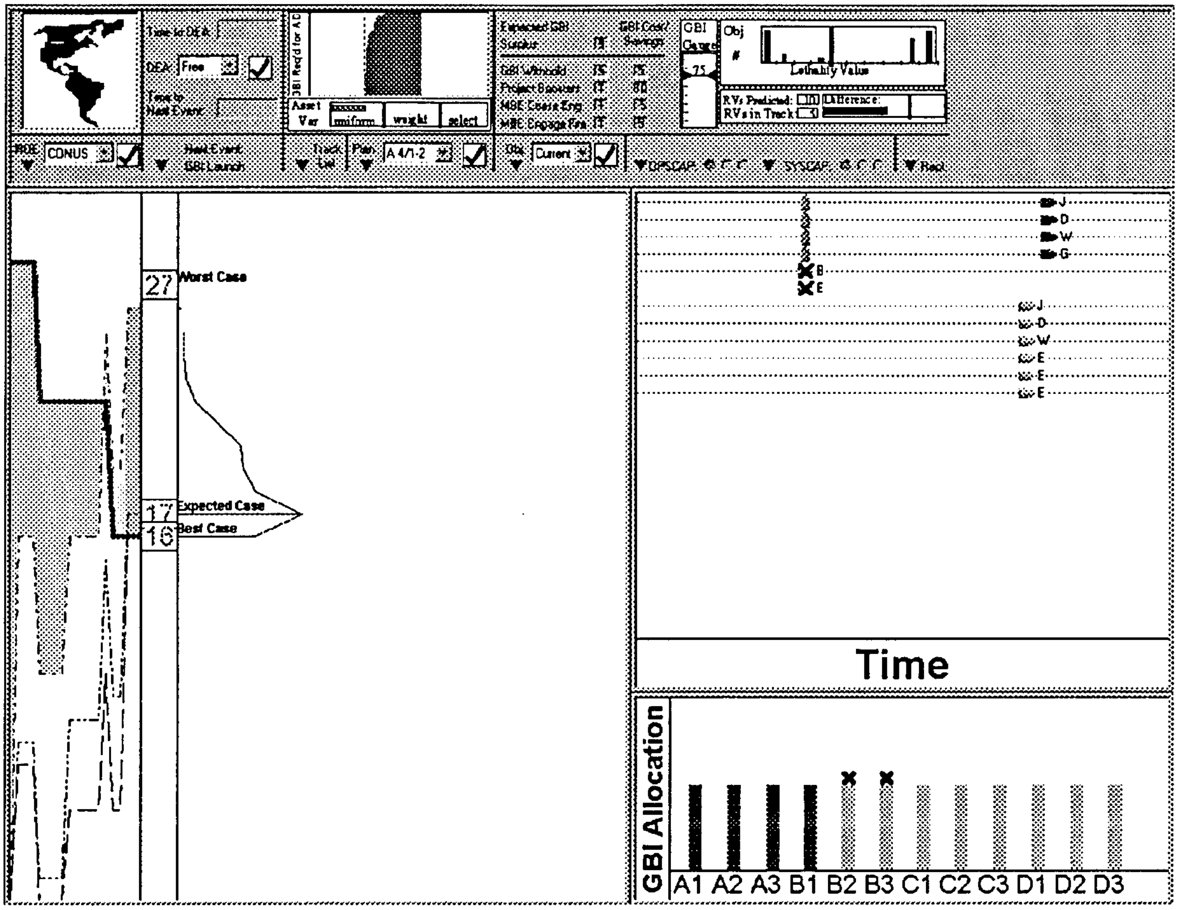


Figure 6. NMD Display Suite: Top horizontal panel and bottom right panel were irrelevant to the task, Risk-Display is the large area on the left, and the Timeline Situation Display is the large area on the right.

Subjects were presented the entire suite and instructed to ignore the top gray rectangular display and the bottom right display, attend to only the Risk-Display (large area on the left) and the Timeline Situation Display (second largest area, middle right). Subjects were told that the Risk-Display was the most important source of information, and were given detailed instructions (Appendix A) on how to interpret it.

2.3.1 Timeline Situation Display

The Timeline Situation Display shows enemy warheads as bullets pointing to the right, moving across the screen as a function of time. A sample Timeline Situation Display is shown in Figure 7.

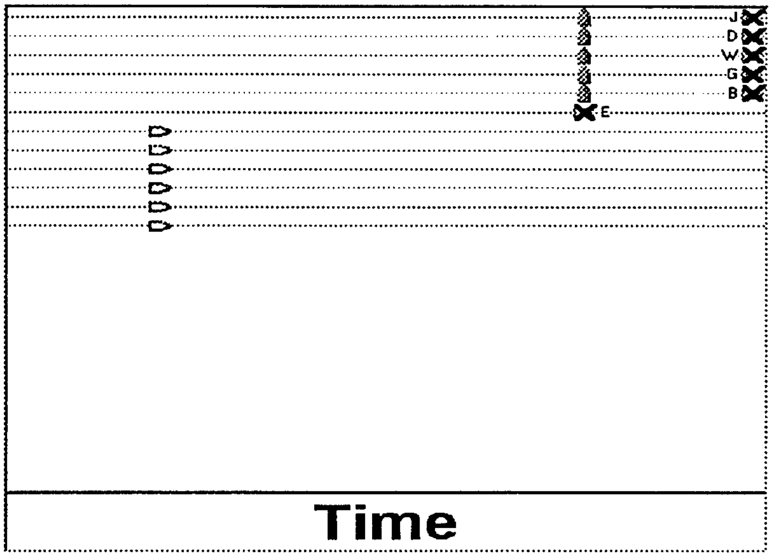


Figure 7. Sample Timeline Situation Display

Bullets pointing up are the GBIs, and they appear on the track of a warhead when the GBI has missed. When an enemy warhead becomes grayed out, with a black X, it has been destroyed. If it reaches the far right side of the screen, it has been destroyed by hitting its intended target, and an upward pointing bullet would show up to indicate a GBI miss. If a warhead becomes grayed out some where along its path, before reaching the right side of the screen, it has been destroyed by one of the GBIs. In Figure 7, the first five enemy warheads (at the top of the display) all hit their targets, and were missed by five GBIs. The sixth warhead was hit by a GBI. In Figure 7, we also see a second launch of six enemy warheads (towards the left of the display), but do not yet know our results against them.

2.3.2 Risk-Display

The Risk-Display is the largest area shown in the NMD display suite, and was introduced to each subject as the most important part of the display. As shown in Figure 1 and in the suite in Figure 6, this display consists of a curve turned on its side which changes shape and position on the screen. The vertical axis indicates the number of GBIs. The heavy black line represents the number of GBIs the system has remaining. For the experimental scenarios, this number always starts out at 28 GBIs. As GBIs are launched at incoming warheads, the number remaining will drop, and the curve will move down the vertical axis. The horizontal axis indicates time, so the curve will move right on the screen as time moves on in a scenario.

The solid thick line and number next to it indicate the total number of GBIs remaining at any given time. In the particular example in Figure 1, six warheads have been identified by the system, but it has not launched any GBIs yet, so all of the original 28 remain in the system. The worst case, expected case and best case probabilities are represented by lines on the curve. The situation where all six warheads leak through the GBI defenses is considered to be the worst case, while the situation where all six warheads are hit by six of the GBIs is considered to be the best case, and the expected case is the median value of the distribution. The 15 additional GBIs in the reserve do not figure into the Risk-Display, or any other part of the display suite

2.3.3 Display Highlighting

Four levels of the highlighting independent variable were employed: best case highlighting, expected case highlighting, worst case highlighting or no highlighting. Highlighting was defined by digitally presenting the number of GBIs associated with the case to be highlighted as well as coloring the corresponding area of the curve in red, and flashing the area slowly by alternating between red and dark red. In the example that was shown in Figure 1, the expected case is highlighted since that number (7) is presented, and the expected case area of the distribution is colored. The display shows that seven GBIs are expected to be needed to defend against the six warheads that have been launched at us. The Risk-Display in Figure 6, shows an example of the no highlighting condition in which no area is highlighted with color and flashing, and the numbers of GBIs required for each case are presented.

2.3.4 Trends

As mentioned earlier, all scenarios included two launches of six enemy warheads each, and four opportunities for the subject to respond as to how many GBIs they want to withdraw from the reserve. Different variations of how many GBIs were hit and how many leaked for each launch (trends) were shown to the subject. Table 1 illustrates how the five levels of the trend variable were defined, in terms of the number of warheads in the air, still coming towards the United States, at each of the feedback (FB) points from the Figure 5. Each column contains two examples of the label at the top of the column.

Bad-Bad	**Bad-Good**	**Expected-Expected**	**Good-Bad**	**Good-Good**
FB1: 4 FB2: 8	FB1: 4 FB2: 4	FB1: 1 FB1: 2	FB1: 0 FB2: 4	FB1: 0 FB2: 0
FB1: 5 FB2: 10	FB1: 4 FB2: 5	FB1: 2 FB2: 4	FB1: 1 FB2: 5	FB1: 1 FB2: 1

Table 1. Levels of the Trend Variable

The five trend variations changed the shape of the Risk-Display curve, since it is computed based on the number of incoming warheads in the air at any given time. For instance, if four warheads from the first launch leaked through the GBI counterattack, they would figure into the curve along with the six warheads from the second launch for a total of ten warheads in the air. Each cell of Table 1 is represented by four scenarios, for a total of forty scenarios per subject. As shown in the above table, some scenarios have the same number of warheads in the air, at the last feedback point, but for differing trends. Examining these scenarios in depth and contrasting judgments on FB2 when preceded by better and worse situations on FB1, will allow us to determine the existence of any order effects.

2.4 Experimental Design

The experiment used a 4x5 factorial design, with both trend and display highlighting varied within subject, for a total of 20 cells. Each cell consisted of two similar versions of the same scenario, for a total of 40 scenarios per subject. The trend independent variable consisted of five levels (bad-good, good-good, bad-bad, good-bad, expected-expected) and the display highlighting independent variable consisted of four levels (worst case highlighted, best case highlighted, expected case highlighted, no highlighting). Both independent variables were blocked and counterbalanced within each block. There were four blocks, which consisted of 40 scenarios each and five subjects saw each block.

2.5 Performance Measures

The subjects response to each question in each scenario (in terms of the number of GBIs to withdraw) was recorded as out main dependent variable, for a total of 160 responses per subject. As mentioned earlier, if a subject exhausted all 15 GBIs allocated in a given scenario, the experiment also recorded the number of GBIs a subject would have chose to withdraw if any were left. A confidence rating was also recorded at the end of each scenario, for a total of 40 confidence ratings per subject. Confidence was defined for the subject as how confident he or she was in the responses they gave for a scenario. This was reported on a scale of one to 10, with 10 being the most confident. Finally, a post-experiment questionnaire (Appendix B) was also administered to each subject.

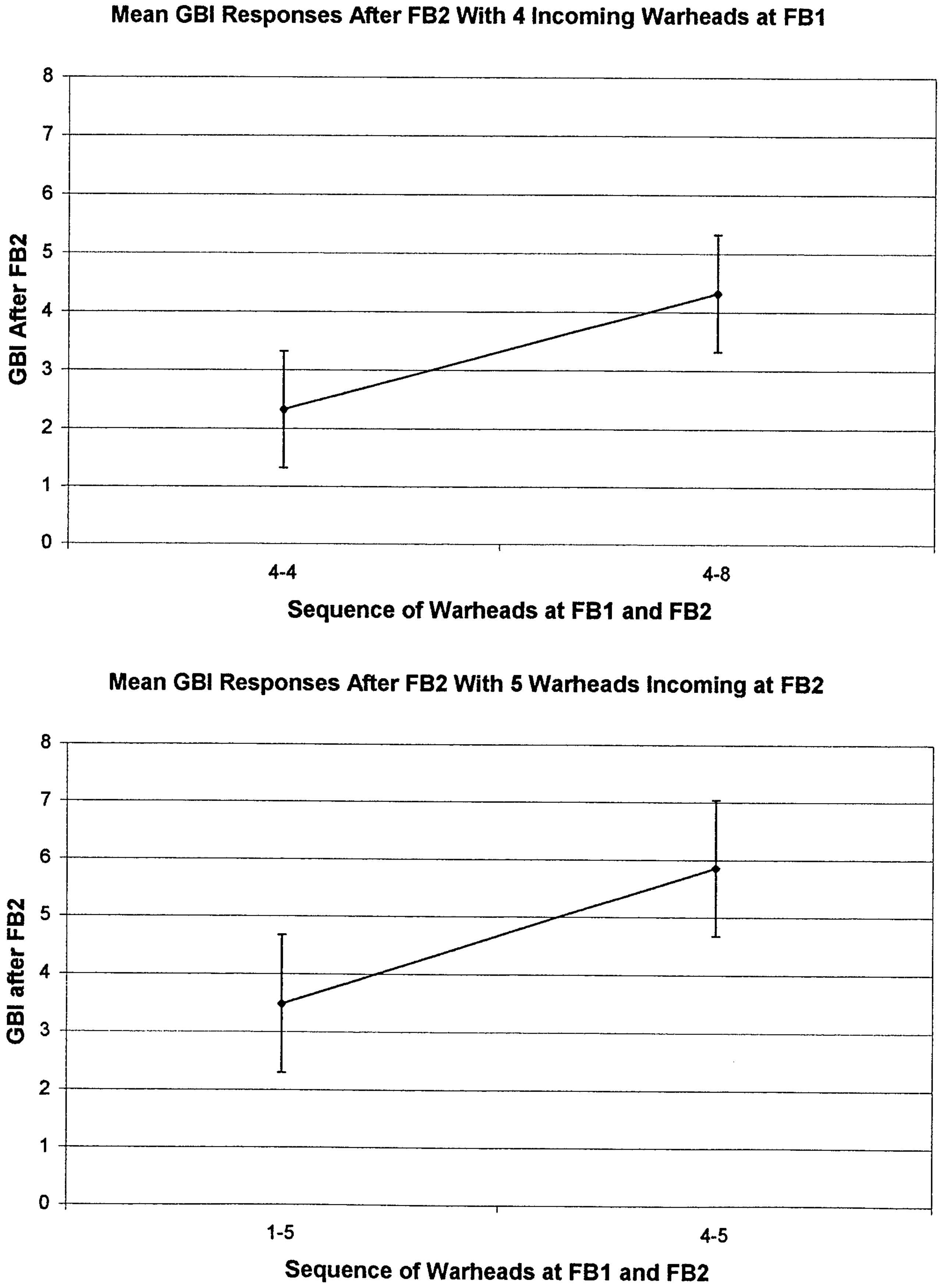
3. RESULTS

All statistical analyses were performed using SAS software from the University of Illinois at Urbana-Champaign. Analysis will be discussed in terms of order effects analysis, trend, display and question effects analysis, analysis of confidence ratings, and analysis of the post-experiment questionnaire.

3.1 Order Effects

As mentioned in the Methods section, to examine possible primacy or recency effects, scenarios which had the same number of warheads in the air at the first or last feedback point (see Figure 5) were analyzed separately, for a total of 16 scenarios per subject. In the primacy analysis, eight scenarios had four warheads in the air at the last feedback point, and eight scenarios had five warheads in the air at the last feedback point. These two groups of eight scenarios were further analyzed separately. Within the group with four warheads at the last feedback point, half of the scenarios had four warheads in the air at the first feedback point (4-4), and half had zero warheads in the air at the first feedback point (0-4). Differing responses for each of these halves would indicate order effects, that is, the response to the same situation at FB2 was influenced differently by the conditions at FB1. Similarly, within the group of five warheads in the air at the last feedback point, half of the scenarios had four warheads in the air at the first feedback point (4-5) and half had one warhead in the air at the last feedback point (1-5).

As shown in Figure 8 (error bars represent standard errors), the mean GBI responses after the second feedback point (question four) with the same number of warheads in the air, differed with the number of warheads in the air at the first feedback point.



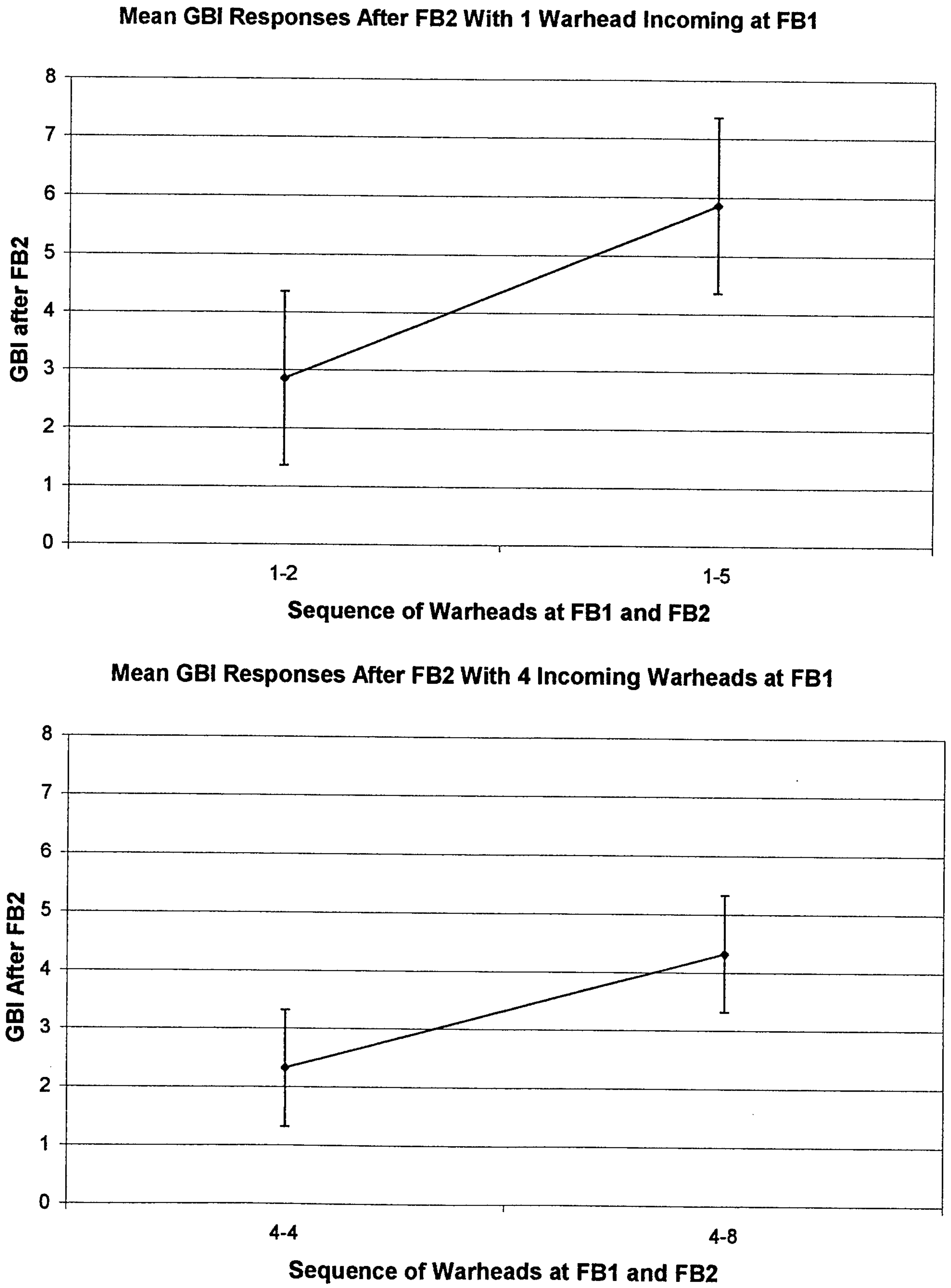
Mean GBI Responses After FB2 With 4 Incoming Warheads at FB1

Mean GBI Responses After FB2 With 5 Warheads Incoming at FB2

Figure 8. Mean GBI Responses

As shown in Figure 8, within the four warheads incoming on FB2 condition, the responses after the last feedback point (question four) were significantly higher when there were zero warheads incoming on FB1 (0-4) than when there were four warheads incoming on FB1 (4-4). Analysis of these scenarios resulted in an F value of 46.05, and a p value of .0001. Similarly, within the five warheads incoming on FB2 conditions, responses were significantly higher when there was one warhead incoming on FB1 (1-5) than when there were four warheads incoming on FB1 (4-5). With the analysis of these scenarios, the F value was 27.97 and the p-value was .0001. These results provide evidence for some primacy in the form of a contrast effect.

To investigate the existence of the more optimal recency effects, scenarios were identified and selectively analyzed which involved the same number of incoming warheads on the first feedback point (FB1), and differing numbers of incoming warheads on the last feedback point (FB2). Eight scenarios were identified which had four enemy warheads incoming at FB1. Of these, half of the scenarios had eight incoming warheads at FB2 (4-8), and half of the scenarios had four incoming warheads at FB2 (4-4). Additionally, eight scenarios were identified which had one warhead incoming at FB1. Of these, half of the scenarios had two incoming warheads at FB2 (1-2) and half of the scenarios had five incoming warheads at FB2 (1-5). As shown in Figure 9, subjects GBI withdrawal responses after the last feedback point (question four) differed with the number of incoming warheads in the air at the second feedback point, indicating some recency.



Mean GBI Responses After FB2 With 1 Warhead Incoming at FB1

Mean GBI Responses After FB2 With 4 Incoming Warheads at FB1

Figure 9. Mean GBI Responses

For the condition with four warheads incoming on FB1, GBI responses for question four were significantly higher for the scenarios with eight incoming warheads on FB2 than the scenarios with four incoming warheads at FB2 (F=25.80, p=.0001). Similarly, for the condition with one incoming warhead on FB1, GBI responses were significantly higher for the scenarios with five incoming warheads on FB2 than the scenarios with two incoming warheads on FB2 (F=65.38, p=.0001). This provides evidence for some recency. From Figures 8 and 9, it is apparent that the effect of primacy was approximately the same size as the effect of recency.

In both of these analyses, it is possible that lower responses on FB2 resulted, not because subjects were more "conservative," but simply because they had expended most of their 15 reserve missiles available to them by that time. In order to determine that the latter factor could not account for the data, a second analysis was run, this time replacing the FB2 response value by the desired value stated verbally by each subject whenever these two values differed. This verbal response was reported any time the subject ran out of reserve GBIs and wished to withdraw more than he or she had remaining. The results of these analyses with the desired GBI responses did not results in different conclusions from those of the actual GBI responses analyses, so only the analysis of the actual GBI responses has been presented.

3.2 Trend, Display and Question Effects

In order to determine the effect of trend (good-bad, good-good, expected-expected, bad-good, bad-bad), display highlighting (worst case, best case, expected case, none), and question (#1, #2, #3, #4) a three-way (5x4x4) Repeated Measures ANOVA was performed with GBI response as the dependent measure. Responses for the two similar scenarios were averaged before the ANOVA, for a total of 20 scenarios per subject. Table 2 presents the results of the ANOVA.

SOURCE	DF	TYPE III SS	MS	F	P
Display	3	1.579	.526	.56	.644
Error (Display)	57	56.615	.941		
Trend	4	779.34	194.835	45.21	.0001
Error (Trend)	76	327.55	4.310		
Question	3	1488.121	496.04	9.60	.0001
Error (Question)	57	2944.148	51.652		
Display x Trend	12	3.898	.328	.72	.730
Error (disp x trend)	228	102.66	.450		
Display x Question	9	21.870	2.430	1.00	.444
Error (disp x question)	171	416.636	2.436		
Trend x Question	12	1256.737	104.728	15.90	.0001
Error (trend x question)	228	1501.995	6.588		
Display x trend x quest	36	46.741	1.299	.93	.589
Error (disp x trend x quest)	684	955.378	1.400		

Table 2. Repeated Measures ANOVA Table

As shown in Figure 10, the question number had a significant effect on the number of GBIs that subjects chose to withdraw (F=9.60, p=.0001), with the number of GBIs increasing with the question number (i.e., as the scenario progressed). Also shown in Figure 10, there was no significant main effect of display highlighting on GBI response, nor was there a significant interaction between display and any other variable.

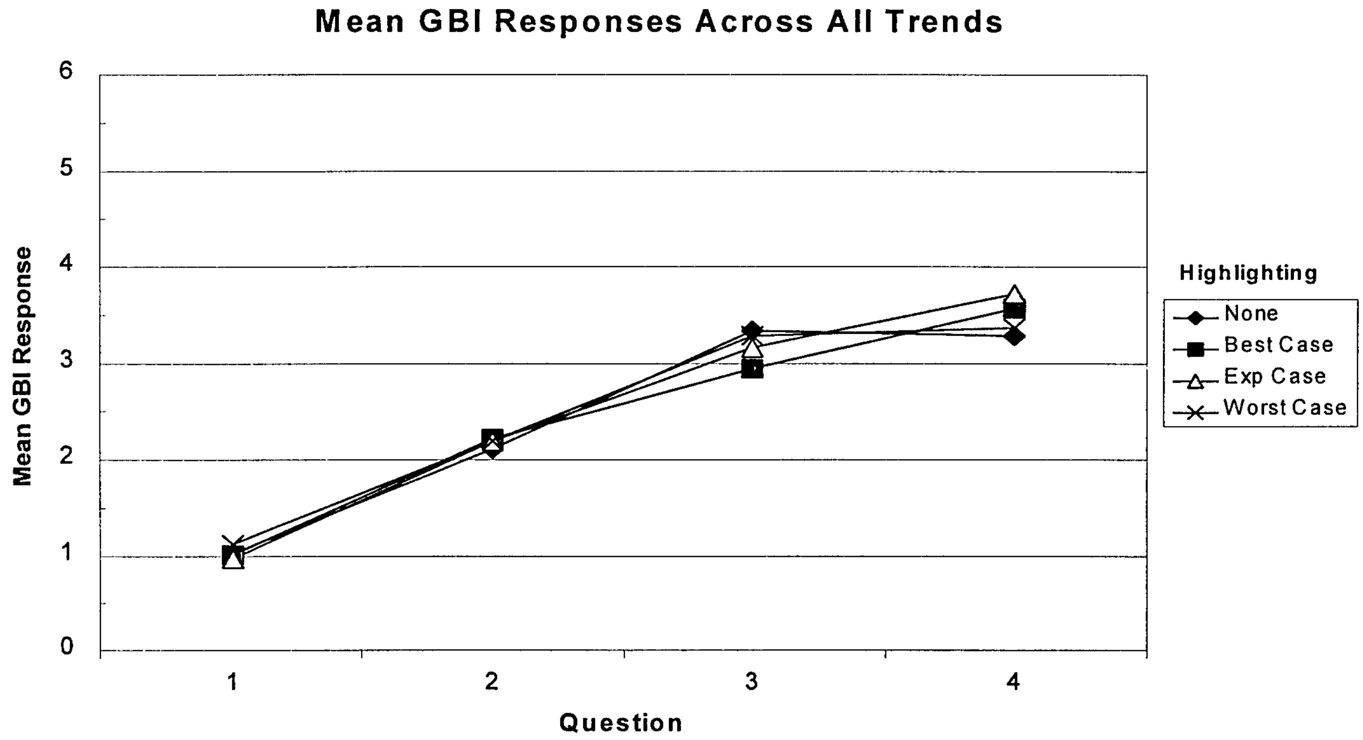


Figure 10. Mean GBI Responses Across Trend

Since trend and question number are related by definition, examining their interaction is more meaningful than the main effects alone. A significant trend by question interaction is shown in Figure 11 (F=15.9, p=.0001) suggesting that the trend changed the effect of the question on GBI responses. As noted in the context of Figure 10, across all trends GBI response tended to increase with the question number. This increase was magnified by trend, such that the GBI response increased the most for the good-bad trend, followed by the bad-bad trend, expected-expected trend, bad-good trend and increased the least with the good-good trend as shown in Figure 11. Such an interaction does indeed suggest that most GBIs were withdrawn when the situation was consistently "bad" (bad-bad) and the least number of GBIs were withdrawn when the situation was consistently "good" (good-good).

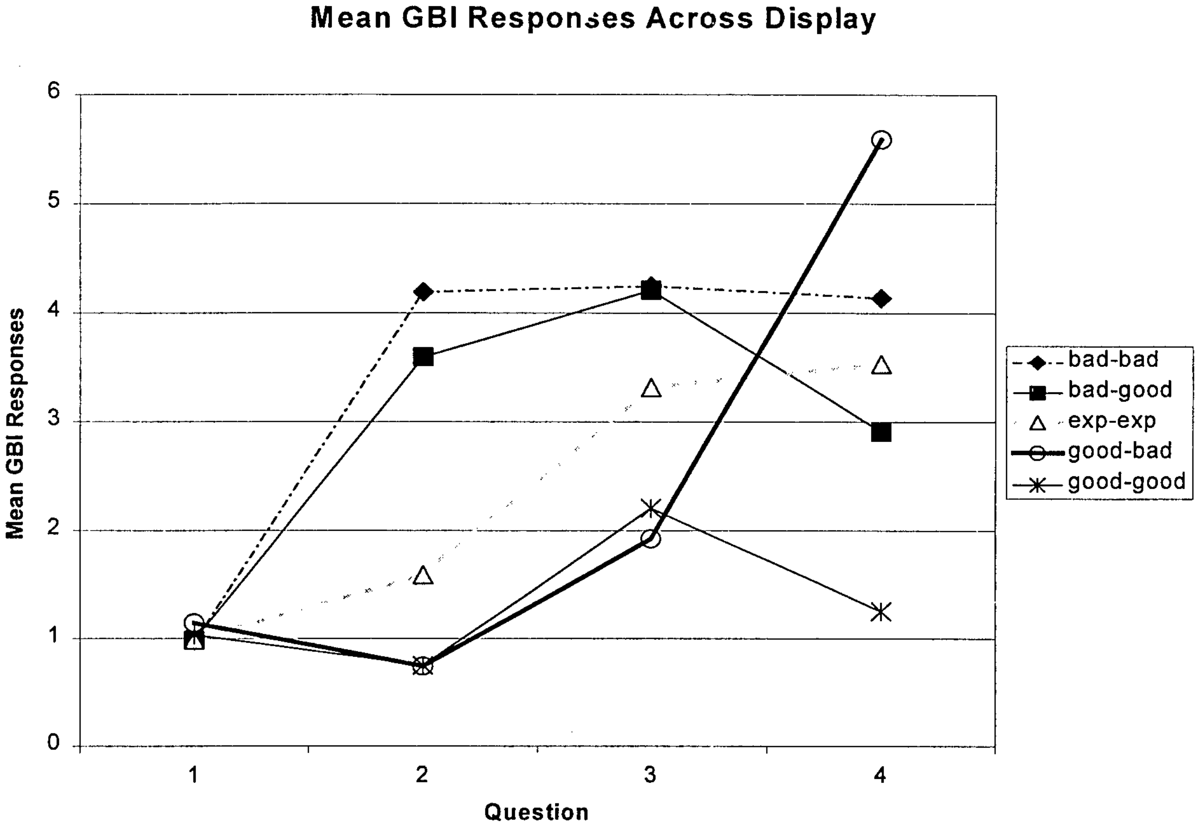


Figure 11. Mean GBI Responses Across Display

More importantly, the trend by question interaction adds evidence for a contrast effect, which can be investigated further by comparing the good-bad trend responses to the bad-bad trend responses. Both trends end up "bad" but start out at different points ("good" or 'bad"), when the first differentiating information becomes available at FB2. In order to visualize the differences, between these two trends, we analyzed them further, separate from the other trends. Figure 12 shows the results of overlaying graphs of the GBI withdrawal responses for the two trends, averaged across display highlighting.

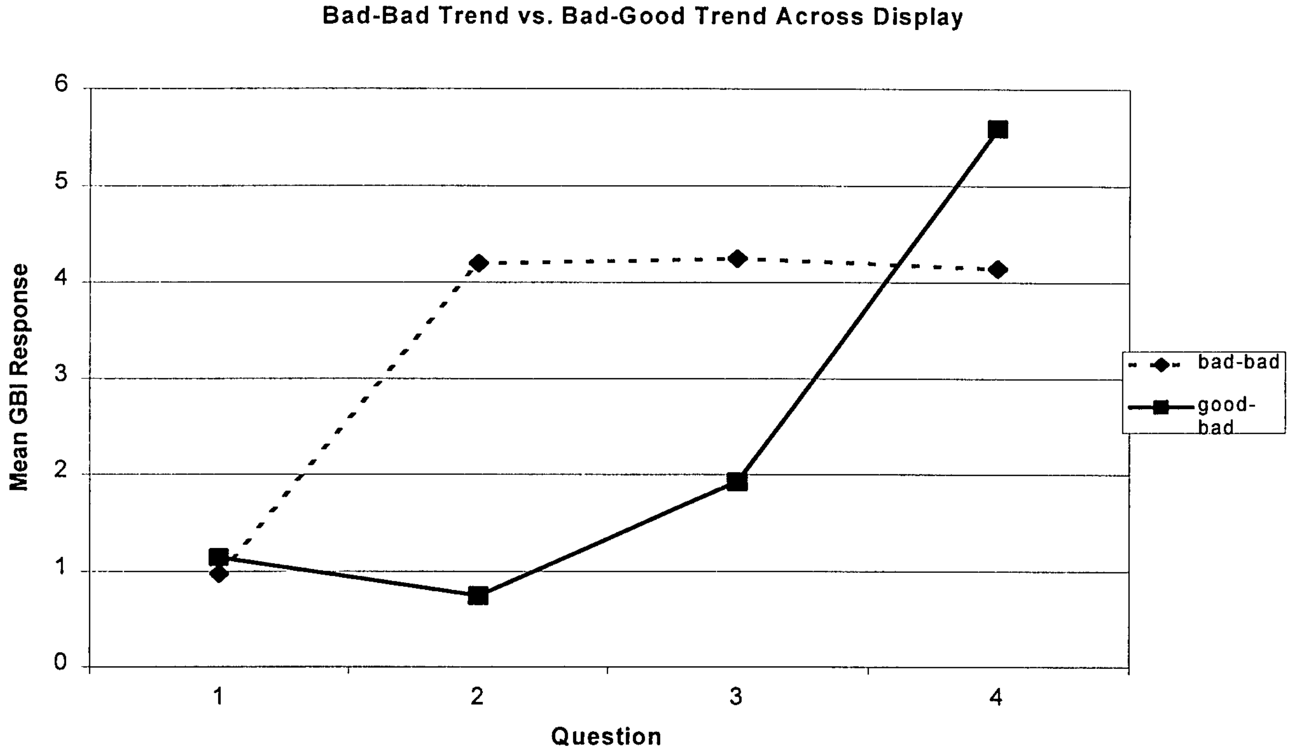


Figure 12. Bad-Bad vs. Good-Bad Trends

As shown in Figure 12, the bad-bad trend responses started out high at question two, but leveled out for the remaining questions. The good-bad trend responses started out low at question two, but showed a very steep increase as the situation became more threatening. Most importantly at question four (when FB2 offered identical information in both conditions), participants withdrew more GBIs if the previous situation had been good than when the previous situation had been bad; here again providing evidence that people are influenced by the trend.

3.3 Confidence Ratings

Confidence ratings were asked of each subject at the end of each scenario on a scale of one to 10 with 10 being the most confident. Only one confidence rating was recorded per scenario (which contained four questions of the subject). Confidence was defined as the confidence in the subjects own decisions to withdraw GBIs from the reserve. In order to examine the confidence ratings of subjects, a two way ANOVA was performed, with trend and display highlighting as independent variables, and confidence rating as the dependent variable. Figure 12 shows the mean confidence ratings for all levels of display highlighting and trend.

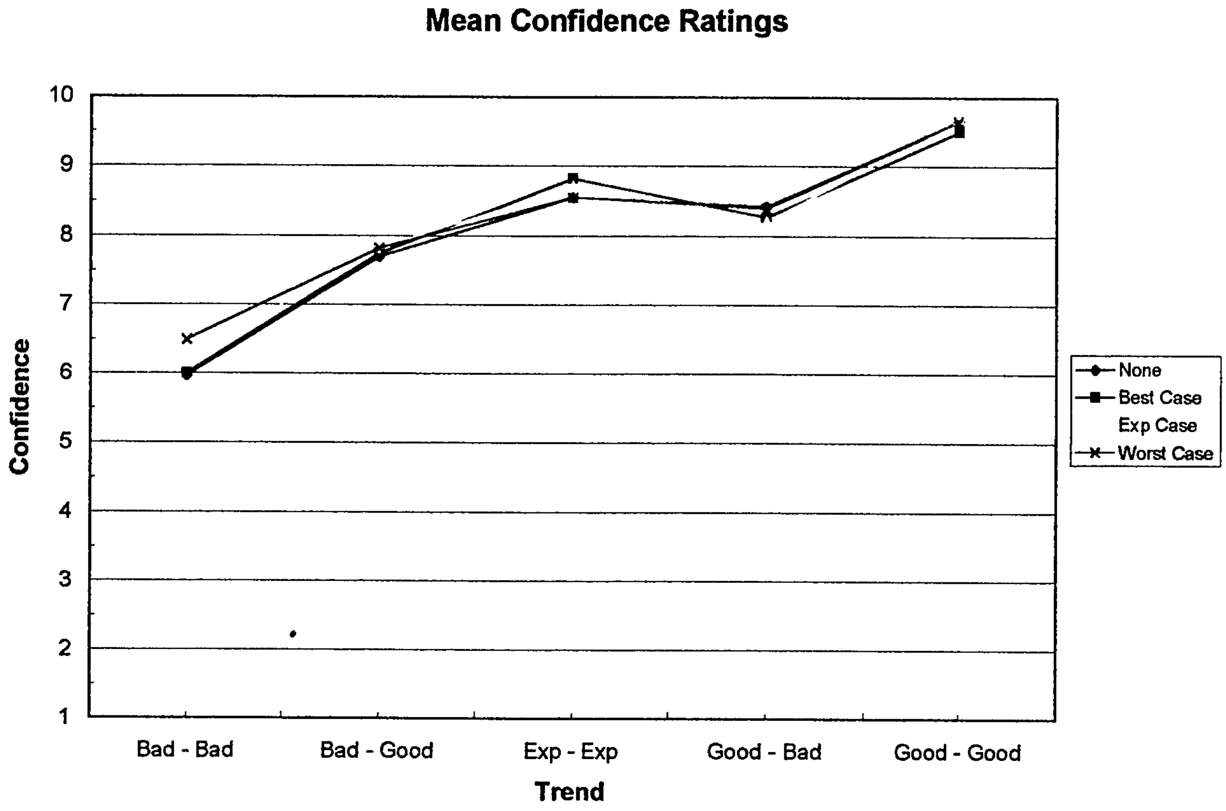


Figure 12. Mean Confidence Ratings

As shown in Figure 12, no significant effect of display type was observed with confidence ratings as the dependent variable, but as with the GBI responses, a significant trend effect was present (F=50.89, =.0001). Table 3 shows the Repeated Measures ANOVA table for confidence ratings. It is apparent that people's judgment was more based upon the confidence that the GBI launching automation was adequately addressing the threat, than upon their confidence that their own judgment to override and withdraw was the correct one.

SOURCE	DF	TYPE III SS	MS	F	P
Display	3	1.477	.492	.47	.701
Error (Display)	57	59.086	1.04		
Trend	4	520.90	130.22	50.89	.0001
Error (Trend)	76	194.48	2.56		
Display x Trend	12	5.364	.447	1.10	.362
Error (display x trend)	228	92.761	.407		

Table 3. Repeated Measures ANOVA for Confidence Ratings

3.4 Post-Experiment Questionnaire

In order to further examine subjects decision making strategies in the NMD task, a post-experiment questionnaire was administered (Appendix B). One question asked what piece of information from the display helped the most in making decisions, best case probability, worst case probability, expected case probability, or the past results of launches. Figure 13 shows the distribution of responses from the 20 subjects:

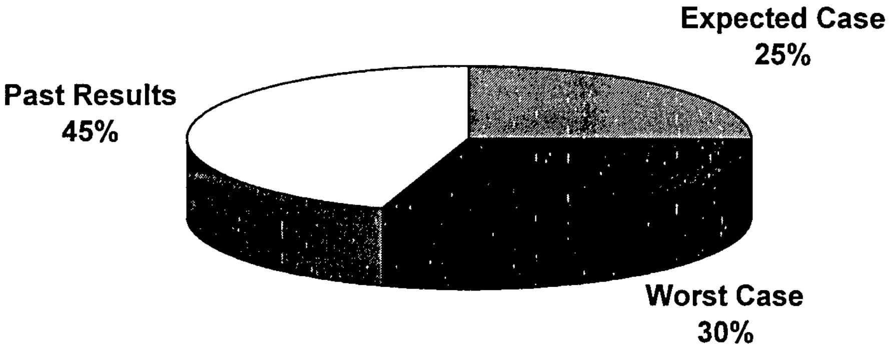


Figure 13. Responses on Most Helpful Piece of Information

Although no display effect was shown through the statistical analysis, 30% of subjects said they used the worst case probability and 25% said they used the expected case probability in making decisions during the task. 45% of subjects said that past results of launches were the most helpful to them, although these results do not affect the probabilities associated with each outcome. This is consistent with the analysis of order effects. Subjects were also asked to rate the ease of interpretation for the Risk-Display and the entire display suite. On a scale of one to seven, with one being the least difficult to interpret and seven being the most difficult to interpret , the mean subject response for the overall NMD computer display was 2.9 with a standard deviation of .09. For the Risk-Display only, the mean subject response was 3.7, with a standard deviation of 1.5. In addition, subjects were asked to provide any further comments about the experiment or the display on the back of the questionnaire form. Only eight subjects provided written comments, and of these, five comments indicated a difficulty in understanding the Risk-Display. These results indicate that the Risk-Display was relatively difficult for subjects to interpret when compared to the display as a whole, which could help to explain the lack of a significant effect due to display highlighting.

4. DISCUSSION

The current experiment was conducted to examine the effects of display highlighting and the order of information presented on operator decision making in the NMD task. We sought to determine if subjects would be influenced by the order effects of primacy and recency when making decisions, and if display highlighting of relevant information could influence decision making and therefore possibly be used to mediate the non-optimal order effects. Contrary to our predictions from examining the relevant literature, no significant effect of display highlighting was observed in our study. In the NMD task, recency, basing decisions on the most recent piece of information, is an optimal decision making strategy since the new information totally replaces the old information. Primacy, basing decisions on the first piece of information presented, would be a non-optimal strategy for the NMD task and result in poor decision making. Recency and primacy effects were both observed with approximately equal effect sizes, indicating some non-optimal tendencies by operators in the NMD task. However, the primacy effect was not one of anchoring, in which later evaluation was pulled in the same direction as previous evidence. Rather the primacy effect was better explained by a contrast effect, in which the current situation is evaluated in terms of the trend from the past to the present. If things are getting better (trend from FB1 to FB2) they are more likely to be judged as "good," whereas if things are getting worse, they are more likely to be judged as "bad." Explanations will now be offered for our findings.

As noted in the results, no effect was observed due to display highlighting, contrary to our predictions from the literature. This lack of a significant effect is possibly due to four factors: first, as reported in the Results section, data from the post-experiment questionnaires indicates that the Risk-Display was relatively difficult for subjects to understand, despite the detailed instructions. Although we emphasized the importance of the Risk-Display in the experiment instructions (Appendix A), we have no guarantee that subjects actually paid attention to the display and used the information it provided to make their decisions. A failure to use information that is difficult to understand when making decisions has been documented in the literature (Bettman, Johnson and Payne, 1991; Johnson, Payne and Bettman, 1988). Subjects could have used only the Timeline Position Display to make their decisions and ignored the Risk-Display, thus not noticing the display highlighting at all. More research must be conducted in order to improve the Risk-Display.

Second, the values of the expected, best and worst case probabilities were highly (although not perfectly) correlated throughout all of the forty scenarios. This correlation would have made it difficult to measure the true effect of display highlighting if one actually existed. Even if subjects did differentially attend to the display highlighting, since the values were correlated, this would have made it more difficult to assess the extent to which the highlighting affected their decisions. More heterogeneous, un-correlated probability values would have possibly have led to a more accurate picture of the influence of display highlighting on operator decision making.

Third, subjects were not informed prior to the block as to which kind of display highlighting they were seeing during the experiment. This information could have aided their decision making by directing their attention before each scenario to the Risk-Display as well as the highlighted information. It also could have helped to make the Risk-Display less difficult to understand by reducing the amount of information to be processed by highlighting a specific part. Making sure subjects were reminded prior to each block about the display highlighting and directing their attention to the Risk-Display could have led to a better representation of the true effect of display highlighting on the NMD task.

Fourth and finally, subject overconfidence in their own abilities or lack of understanding of the automation may have caused them to ignore (or under-trust) the Risk-Display and thus not pay attention to display highlighting when making decisions. This idea is supported by the literature concerning human overconfidence during interaction with automated systems. For instance, Riley (1996) and Liu, Feld and Wickens (1993), found that subjects tended to show this overconfidence by relying on their own performance even when the performance of the automated system was superior. Additionally, Kleinmuntz (1990) evaluated the use of decision-making aids for subjects and found that people viewed the decision aids as inaccurate and relied on their own abilities instead of using the information provided in the decision aid. For whatever reason, if subjects did not attend to the Risk-Display during the experiment, highlighting pieces of information within this display would have not influenced their decisions.

Main effects of question and trend were observed as well as a significant interaction between the two variables. The first general finding here is that subjects chose to withdraw more GBIs as time went on in the scenarios, regardless of the trend. This can be considered a rational response because more incoming warheads would be in the air as time went on, presenting more of a threat. The next finding is that as the situation got worse (number of incoming warheads increased), GBI withdrawal responses tended to increase. This can also be considered a rational behavior since worsening situations would present more of a threat and would require more GBIs for defense.

While the analysis of trend and question variables thus indicated some degree of rational and optimal subject behavior for the NMD task, a more detailed analysis did reveal non-optimal order effects. Sequential effects were examined for primacy and recency given that recency was optimal for the NMD task. While there *was* indeed good evidence for recency, there was also good evidence for the less optimal primacy effect in the form of a contrast effect, in that response on FB2 was influenced by the *trend* from FB1 to FB2. A recency effect was uncovered by selectively analyzing scenarios with the same number of warheads incoming at the first feedback point, and differing numbers of warheads incoming at the second feedback point. As we mentioned, recency is the preferred tactic in making decisions in the NMD environment, because new information completely updates the old information and becomes the most accurate and relevant information for the task. However, subjects did not show perfect recency. If subjects showed perfect recency, then their judgments at FB2 should be totally unaffected by what happened at FB1. This is clearly not the case with our data.

A contrast effect was uncovered by examining scenarios with the same number of warheads incoming at the second feedback point, and differing numbers of warheads incoming at the first feedback point. In optimal behavior, without order effects, there should be no differences in responses for these scenarios. Further evidence for recency was found in comparing the good-bad trend to the bad-bad trend as shown in Figure 12. Even though the trends both end up bad (with a large number of missiles in the air), subjects overcompensated with their GBI responses for the good-bad trend. This indicates that subject decision making was influenced by the worsening of a situation not only by the present point, which can be considered a contrast or trend effect.

These order effects could help to explain the "trigger happy" tendency as reported in preliminary analysis of the NMD task (M. Barnes, personal communication, 1999), and has been reported in the literature. For instance, primacy was also uncovered in a complex decision making task by Tolcott, Marvin and Bresoick (1989). The existence of a recency effect supports the findings of Hogarth and Einhorn (1992), which stated that for a complex decision making task, where judgments are required several times during the sequence of information presentation, the recency effect is observed. While recency is a more optimal form of decision making in the NMD task, non-optimal primacy was also observed, and should be taken into consideration when designing automated systems. Although recency is considered to be optimal for the NMD task, trend perception may also be considered optimal under certain conditions. For instance, if a worsening trend is diagnostic of system failure, then trend perception and increased pessimism would be optimal behaviors. In a system with lags, trend perception would also be an optimal form of operator behavior in order to anticipate future system responses.

This study indicates at least two future directions for experimental research concerning the NMD system mentioned previously. Since the display highlighting variable alone was not an effective means of altering subject decision making, experimental changes should take place to determine if there truly is an effect of display highlighting. First, in order to ensure that the highlighting is noticed, the Timeline Position Display could be deleted from the NMD display suite, forcing subjects to attend to the Risk-Display to obtain the information they need. Second, the values of the different cases (best case, worst case, expected case) could be altered to ensure that they were not highly correlated. We expect these experimental changes to lead to a significant display effect as predicted in the literature (Payne, 1980).

REFERENCES

Adelman, L., and Bresnick, T. (1992). Examining the effect of information sequence on Patriot Air Defense Officers' judgments. *Organizational Behavior and Human Decision Processes*, 53, 204-228.

Andre, A., and Cutler. (1998). Displaying uncertainty in navigation systems. *Proceedings of the 42nd Annual Meeting of the Human Factors and Ergonomics Society.* Santa Monica, CA: Human Factors and Ergonomics Society.

Bettman, J.R., Johnson, E.J., and Payne, J.W. (1991). Consumer decision making. In T.S. Robertson and H.S. Kassarjin (eds.), *Handbook of consumer behavior* (pp. 50-84). New York: Prentice Hall.

Bornstein, B.H., and Zickafoose, D.J. (1999). I know I know it, I saw it: The stability of the confidence-accuracy relationship across domains. *Journal of Experimental Psychology: Applied,* 5(1), 76-88.

Carroll, J.S. (1980). Analyzing behavior: The magician's audience. In T.S. Wallsten (ed.), *Cognitive processes in choice and decision making.* Hillsdale, NJ: Erlbaum.

Combs, B., and Slovic, P. (1979). Newspaper coverage of causes of death. *Journalism Quarterly*, 56(4), 837-843; 849.

Edwards, W. (1962). Dynamic decision theory and probabilistic information processing. *Human Factors*, 4, 59-73.

Fischoff, B., Slovic, P., and Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 3, 552-564.

Gempler, K.S., and Wickens, C.D. (1998). *Display of predictor reliability on a cockpit display of traffic information.* University of Illinois Institute of Aviation Final Technical Report (ARL-98-6/ROCKWELL-98-1). Savoy, IL: Aviation Research Lab.

Hogarth, R.M., and Einhorn, H.J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24, 1-55.

Johnson, E.J., Payne, J.W., and Bettman, J.R. (1988). Information displays and preference reversals. *Organizational Behavior and Human Decision Processes*, 42, 1-21.

Kaplan, C.A., and Simon, H.A. (1990). In search of insight. *Cognitive Psychology*, 22, 374-419.

Kirschenbaum, S.S., and Arruda, J.E. (1994) Effects of graphic and verbal probability information on command decision making. *Human Factors*, 36(3), 406-418.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin*, 107(3), 296-310.

Liu, Y., Fuld, R., and Wickens, C.D. (1993). Monitoring behavior in manual and automated scheduling systems. *International Journal of Man-Machine Studies*, 39, 1015-1029.

MacGregor, D., and Slovic, P. (1986). Graphic representation of judgmental information. *Human-Computer Interaction*, 2, 179-200.

McFadden, S.M, Giesbrecht, B.C., and Gula, C.A. (1998) Use of automatic tracker as function of its reliability. *Ergonomics*. 41(4), 512-530.

McNeil, B.J., Pauker, S.G., Sox, H.C., Jr., and Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England Journal of Medicine,* 306, 1259-1262.

Merlo, J.L., Wickens, C.D., and Yeh, M. (1999). *Effect of Reliability on Cue Effectiveness and Display Signaling.* University of Illinois Institute of Aviation Final Technical Report (ARL-99-4/FED-LAB-99-3). Savoy, IL: Aviation Research Lab.

Montgomery, D.A. (1999). Human sensitivity to variability information in detection decisions. *Human Factors*, 41(1). 90-105.

Muir, B.M, and Moray, N. (1996). Trust in Automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3), 429-460.

Nygren, T. E. (1997). Framing of task performance strategies: Effects on performance in a multiattribute dynamic decision making environment. *Human Factors* 39(3), 425-437.

Parasuraman, R. (1986). Vigilance, monitoring, and search. In K. Boff, L. Kaufman, and J. Thomas (Eds.), *Trends in ergonomics/human factors II* (pp. 59-66). Amsterdam: North-Holland.

Parasuraman, R., and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230-253.

Parasuraman, R., Mouloua, M., Molloy, R., and Hilburn, B. (1996). Monitoring of automated systems. In R. Parasuraman and M. Mouloua (Eds.), *Automation and human performance: Theory and application* (pp. 91-115). Mahwah, NJ: Erlbaum.

Payne, J.W. (1980). Information processing theory: Some concepts and methods applied to decision research. In T.S. Wallsten (ed.), *Cognitive processes in choice and decision behavior*. Hillsdale, NJ: Erlbaum.

Peterson, C.R., and Beach, L.R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, 68, 29-46.

Pitz, G.F. (1980). The very guide of life: The use of probabilistic information for making decisions. In T.S. Wallsten (ed.), *Cognitive processes in choice and decision behavior.* Hillsdale, NJ: Erlbaum

Puto, C.P., Patton, W.E., III, and King, R.H. (1985). Risk handling strategies in industrial vendor selection decisions. *Journal of Marketing*, 49, 89-98.

Riley, V. (1996). Operator reliance on automation: Theory and data. In R. Parasuraman and M. Mouloua (Eds.), *Automation and human performance: Theory and application* (pp. 19-35). Mahwah, NJ: Erlbaum.

Sarter, N.B., and Woods, D.D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors*, 37(1), 5-19.

Schipper, L.M., and Doherty, M.E. (1983). *Decision making and information processing under various uncertainty conditions.* Brooks Air Force Base, TX.: Air Force Human Resources Laboratory, Air Force Systems Command. Series: AFHRL-TR. 83-19

Schum, D. (1975). The weighting of testimony of judicial proceedings from sources having reduced credibility. *Human Factors*, 17, 172-203.

Schurr, P.H. (1987). Effects of gain and loss decision frames on risky purchase negotiations. *Journal of Applied Psychology*, 72(3), 351-358.

Schwartz, D.R., and Howell, W.C. (1985). Optional stopping performance under graphic and numeric CRT formatting. *Human Factors*, 23, 541-550.

Slovic, P. (1987). Facts vs. fears: Understanding perceived risk. In F. Farley and C.H. Null (eds.), *Using psychological science: Making the public case* (pp. 57-68). Washington, D.C.: The Federation of Behavioral, Psychological and Cognitive Sciences.

Stone, E.R., Yates, J.F., and Parker, A.M. (1997). Effects of numerical and graphical displays on professed risk-taking behavior. *Journal of Experimental Psychology: Applied*, 3(4), 243-256.

Tolcott, M.A., Marvin, F.F., and Bresoick, T.A. (1989). *The confirmation bias in military situation assessment.* Reston, VA: Decision Science Consortium.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.

Tversky, A., and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211, 453-458.

Waganaar, W.A., and Sagaria, S.D. (1975). Misperception of exponential growth. *Perception and Psychophysics*, 18, 416-422.

Wallsten, T.S., and Barton, C. (1982). Processing probabilistic multidimensional information for decisions. *Journal of Experimental Psychology: Learning, Memory and Cognition* 8, 361-384.

Wells, G.L., Lindsay, R.C., and Ferguson, T.I. (1979). Accuracy, confidence and juror perceptions in eyewitness testimony. *Joutnal of Applied Psychology*, 64, 440-448.

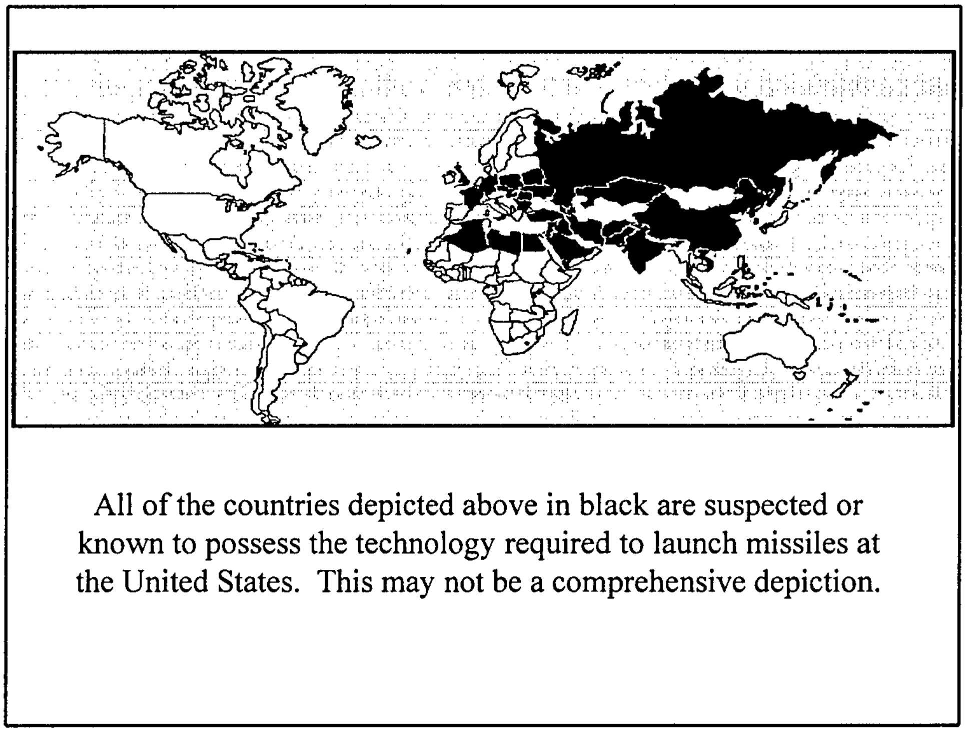
Wickens, C.D. (1992). *Engineering Psychology and Human Performance* (2nd ed.). New York: HarperCollins.

Wickens, C.D., and Hollands, J. (in press, 1999). *Engineering Psychology and Human Performance (3rd Ed.)* New York: Prentice Hall.

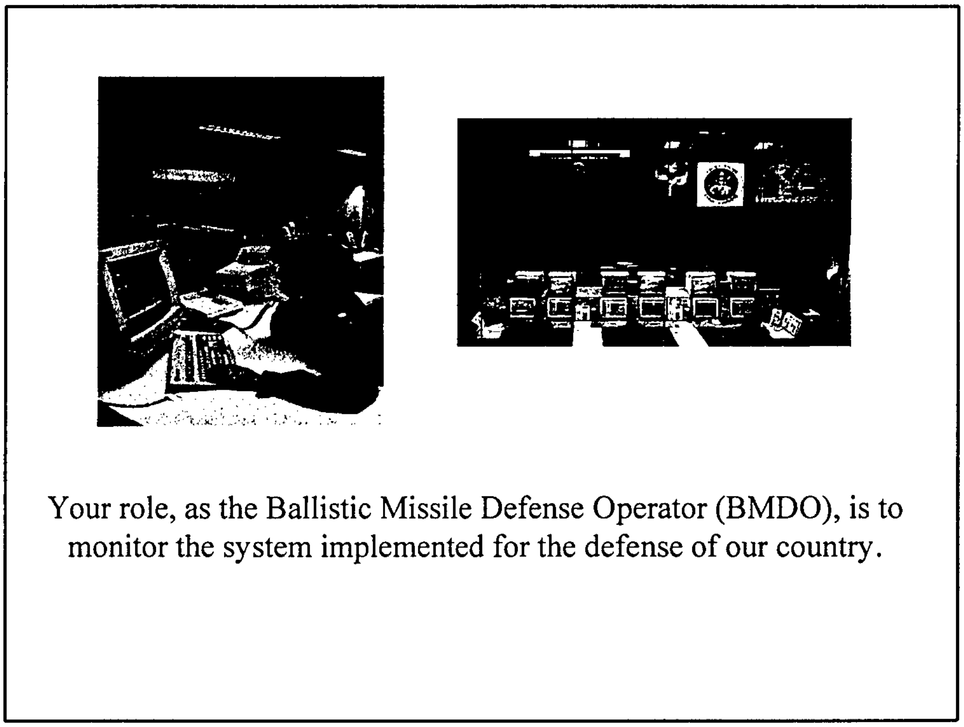
Wickens, C.D., and Kessel, C. (1979). The effect of participatory mode and task workload on the detection of dynamic system failures. *IEEE Transactions on Systems, Man, and Cybernetics,* 13, 21-31.

Wickens, C.D., and Morphew, E. (1997). *Predictive features of a cockpit traffic display: A workload assessment.* University of Illinois Institute of Aviation Final Technical Report (ARL-97-6/NASA-97-3). Savoy, IL: Aviation Research Lab.

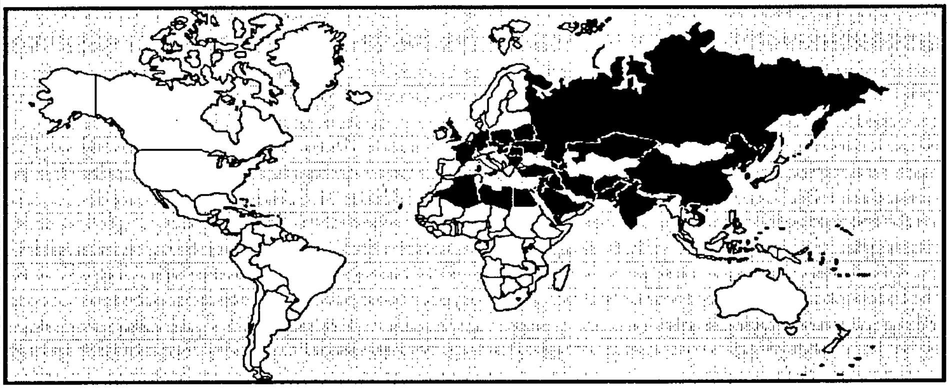
Wickens, C.D., Gordon, S.E., and Liu, Y. (1998). *An introduction to human factors engineering.* New York: Wesley Longman, Inc.



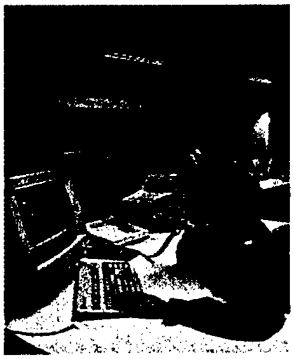
All of the countries depicted above in black are suspected or known to possess the technology required to launch missiles at the United States. This may not be a comprehensive depiction.

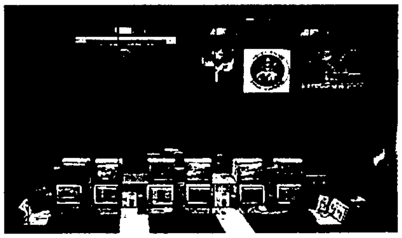


Your role, as the Ballistic Missile Defense Operator (BMDO), is to monitor the system implemented for the defense of our country.



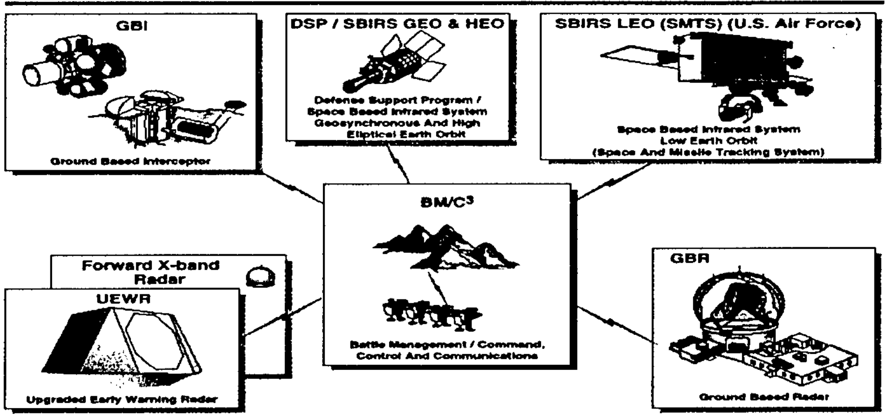
All of the countries depicted above in black are suspected or known to possess the technology required to launch missiles at the United States. This may not be a comprehensive depiction.



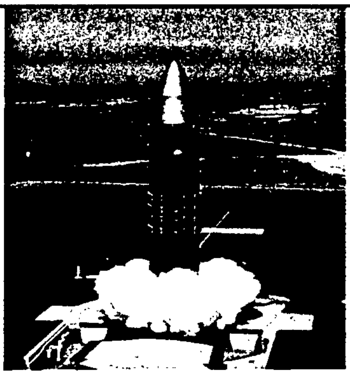


Your role, as the Ballistic Missile Defense Operator (BMDO), is to monitor the system implemented for the defense of our country.

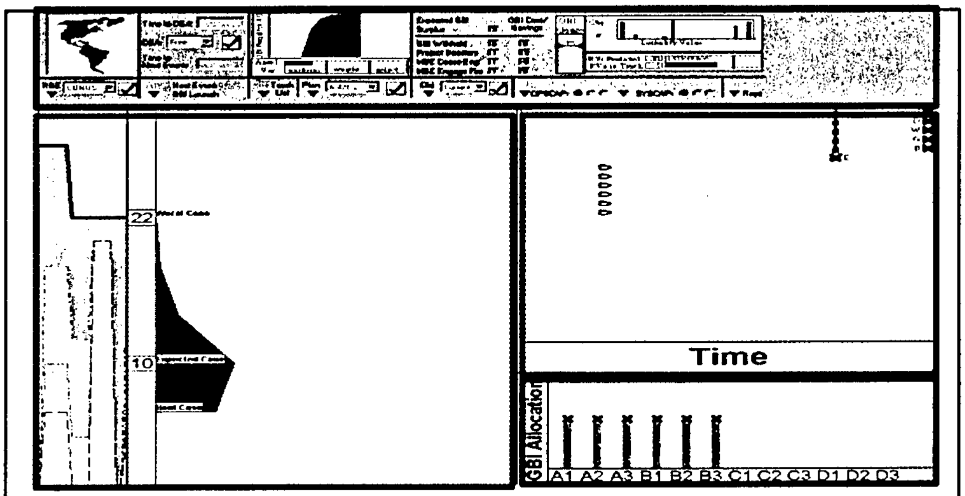
NMD SYSTEM ELEMENTS


GBI
Ground Based Interceptor
DSP / SBIRS GEO & HEO
Defense Support Program / Space Based Infrared System Geosynchronous And High Elliptical Earth Orbit
SBIRS LEO (SMTS) (U.S. Air Force)
Space Based Infrared System Low Earth Orbit (Space And Missile Tracking System)
BM/C3
Battle Management / Command, Control And Communications
Forward X-band Radar
UEWR
Upgraded Early Warning Radar
GBR
Ground Based Radar

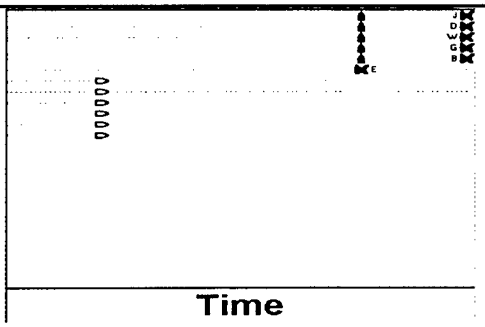
In general, the National Missile Defense (NMD) system consists of radar and satellite systems to track missile launches, a ground based interceptor system to target and destroy incoming missiles, and you, the BMDO.



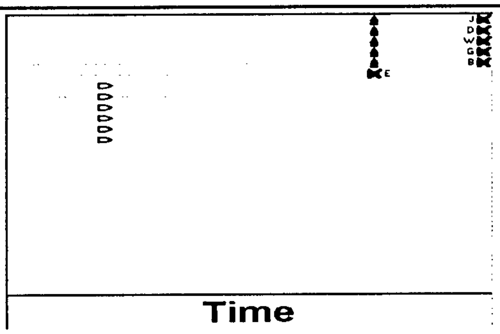
As missiles are launched against us, the automated NMD system tracks them and launches Ground Based Interceptors (GBIs) at each incoming warhead. These GBIs are limited in number and must be conserved. A reserve supply of GBIs does exist for you to use, if necessary. As the monitor of the system, your job is to act as a final safeguard, and to determine if the actions taken by the automation appear to be correct, and in particular, if it appears necessary to withdraw GBIs from the reserve.



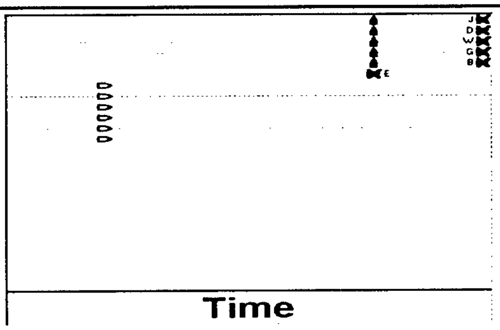
The NMD display screen depicted above, gives you all of the pertinent information concerning the actions of the automated system that you are to monitor. The parts of the display outlined in black contain information for other operators, and is of no use to you. The rest of the display contains two parts: the Risk Display (highlighted in green), and the Current Missile Display (purple).



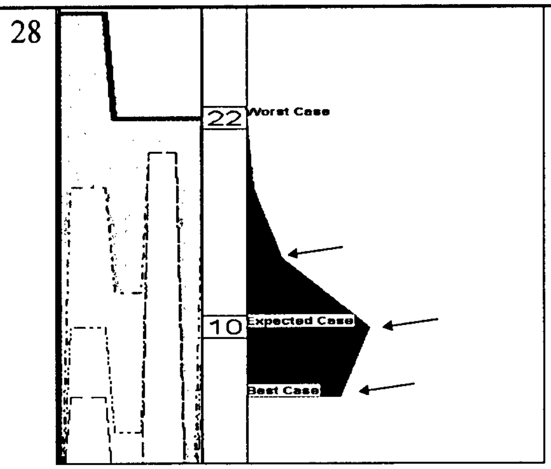
As warheads are launched at the US, the Current Missile Display shows each warhead moving across the screen as a function of time, from launch toward the destination (right side of the screen). Active warheads are shown in red. When the automation suspects a warhead exists, it shows up as an outlined red bullet (like above). When the automation confirms that suspected warheads are actual warheads, the bullet will become solid red. When the automation determines where the warhead is going, a letter designator will appear by the warhead. When the automation launches our GBIs at a warhead, the targeted warhead flashes yellow and red.



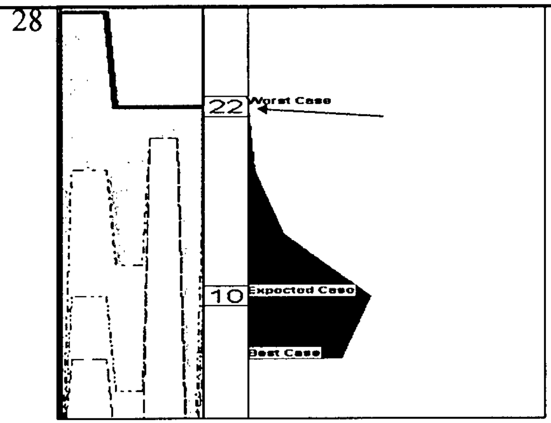
Destroyed warheads are shown in gray with an X. There are two ways for a warhead to be destroyed: by hitting its target, or by being hit by a GBI. If a GBI destroys a warhead, it will appear as warhead E above. If a GBI misses a warhead, the GBI will show up as a light blue bullet, and the warhead will continue to travel across the screen until it is destroyed by another GBI or hits its intended destination and turns gray (like warheads J,D,W,G & B above).



As incoming warheads travel across the earth (and your screen) we will have multiple opportunities to destroy them. However, beyond a certain point, it becomes too late, and the automation will have no more chances to shoot down a warhead with GBIs before it hits its final target. Furthermore, our GBIs are not perfect. There is approximately a 20% chance that a GBI will miss the warhead it was targeted to hit. Since GBIs can miss, there is additional risk involved that is shown to the operator in the form of the Risk Display.



The Risk Display is the most important display feature. This shows the total number of GBIs available at any given point (22 in the example above), as well as a distribution of outcome possibilities that can be represented by three probabilities. The three dividing lines in the Risk Display (marked with arrows above) represent the probabilities, or chance, that each case could occur.

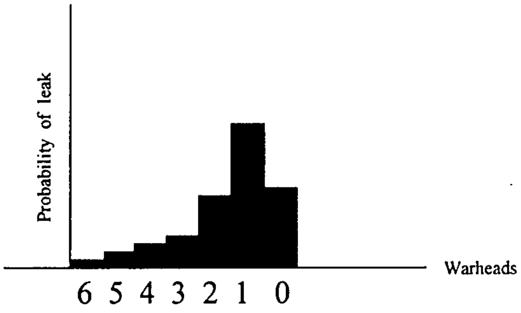


You have 28 GBIs in the system at the beginning of each scenario. Each time enemy warheads are launched, the automation fires 1 GBI at each warhead, decreasing the total number of GBIs available. In the above example, 6 warheads have been launched at us, so the automation fired 6 GBIs back, decreasing the total from 28 to 22 (marked with arrow). There are also 15 GBIs set aside in the reserves, some or all of which you can choose to withdraw when you are given the opportunity to do so.

The Risk Display will now be discussed in more detail. Since this is the most important source of information for you, it is important that you understand this display thoroughly.

As we mentioned, our GBIs are not perfect- there is only an 80% probability that a GBI will destroy the warhead it was targeted at. This leaves a 20% probability that the GBI will miss, and the warhead will survive, or "leak" through our defenses. If there are 2 warheads fired against us, there is only a 64% probability (80% squared) that both of them will be destroyed by our GBIs, which leaves a 32% probability that 1 will be destroyed and 1 will leak, and a 4% probability that both will leak. Destroying both warheads in this case, is considered to be the "best case," while missing both warheads is called the "worst case."

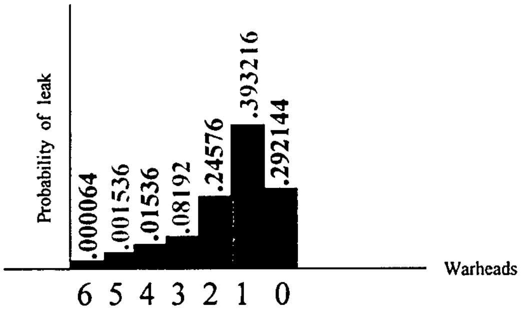
If several warheads are fired at us, there are many different possibilities of how many leak and how many are destroyed. The calculation of the probabilities associated with each possibility becomes more complex, so we will represent these probabilities as a graphical distribution. Consider the situation that 6 warheads are fired against us. The following distribution exists:



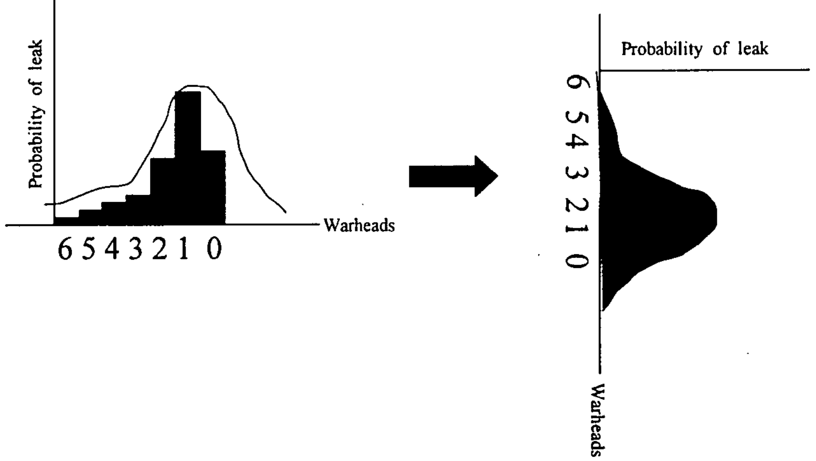
Shown on the graph below, are the probabilities of leaking associated with each event, for the case that 6 warheads are launched against us. They are calculated with the binomial formula. For instance,

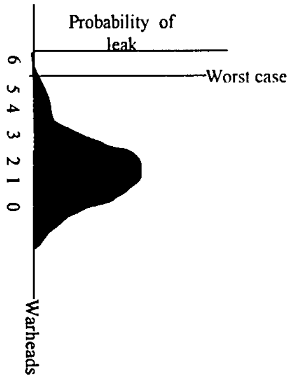
The probability that 5 out of 6 warheads leak=

$$\left(\frac{6!}{(5!)(1!)}\right).2^5.8^1 = .001536$$

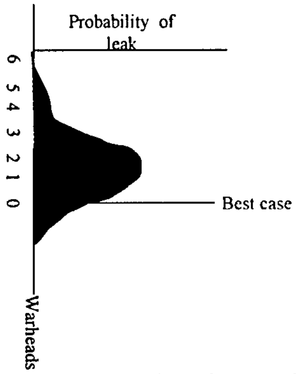


The same type of distribution will be shown on your Risk Display, with a curve fit to it, and turned on its side:

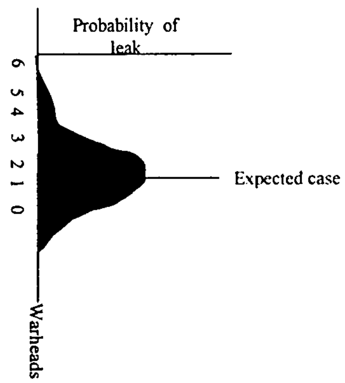




Lets consider this example further. There are three important parts to this distribution. The worst case, is the case when all six warheads leak, and more GBIs must be launched. This is a very low probability that is calculated by the automated system. Remember that when only 2 warheads were fired at us, the probability that both warheads will leak was only 4% , so the probability that 6 warheads will leak when 6 are fired at us will be much smaller.



The best case occurs when all 6 warheads are destroyed by our GBIs, so no more GBIs will be needed to deal with them. This probability is also calculated by the automation. Remember with 2 warheads launched, it was 64%, so with 6 warheads, this probability will also be lower.

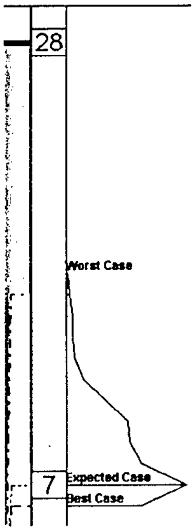


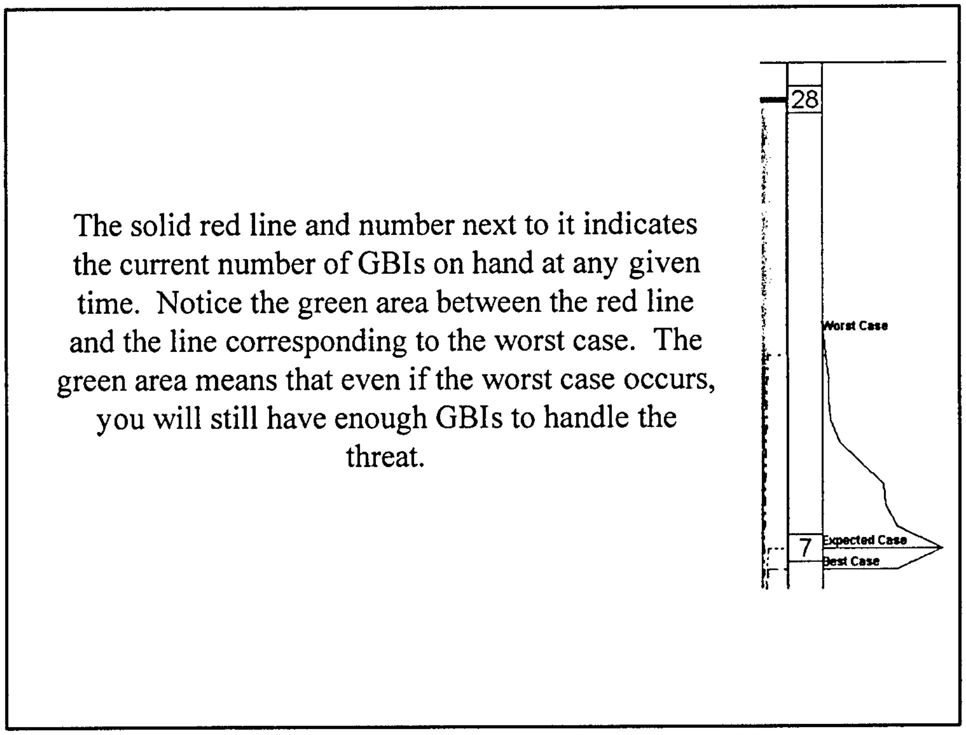
Finally, the expected case can be thought of as a long term average of how many warheads will be destroyed. It is also estimated by the automation, and is at the center of our distribution.

Throughout the experiment, you need to examine the distributions shown and determine if the level of risk they show demands the launching of more GBIs.

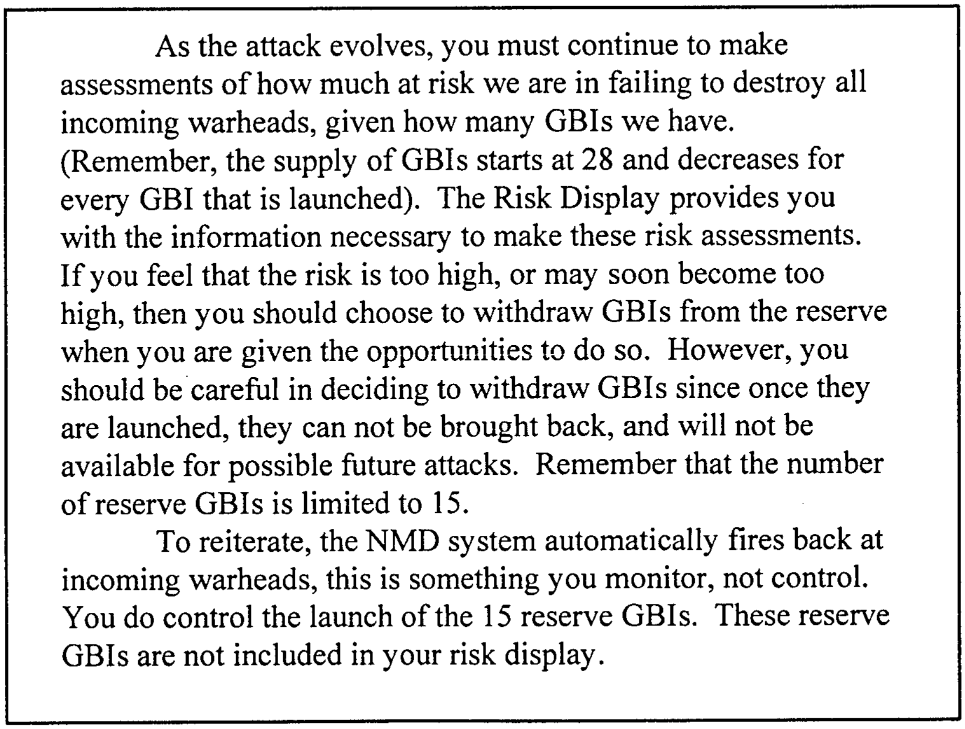
Remember, to further help you, the Risk Display also shows the total number of GBIs you have on hand on the left side of the distribution (not including the 15 in reserve).

The example on the right of this screen is one example of a Risk Display. Shown is the distribution (outlined in blue), as described earlier, which moves to the right of the screen as time goes on. It will also move up and down the screen as GBIs are launched, and will change its shape as different numbers of un-destroyed warheads remain traveling toward us. Notice the three dashed lines to the left of the distribution correspond to the worst, expected and best cases.



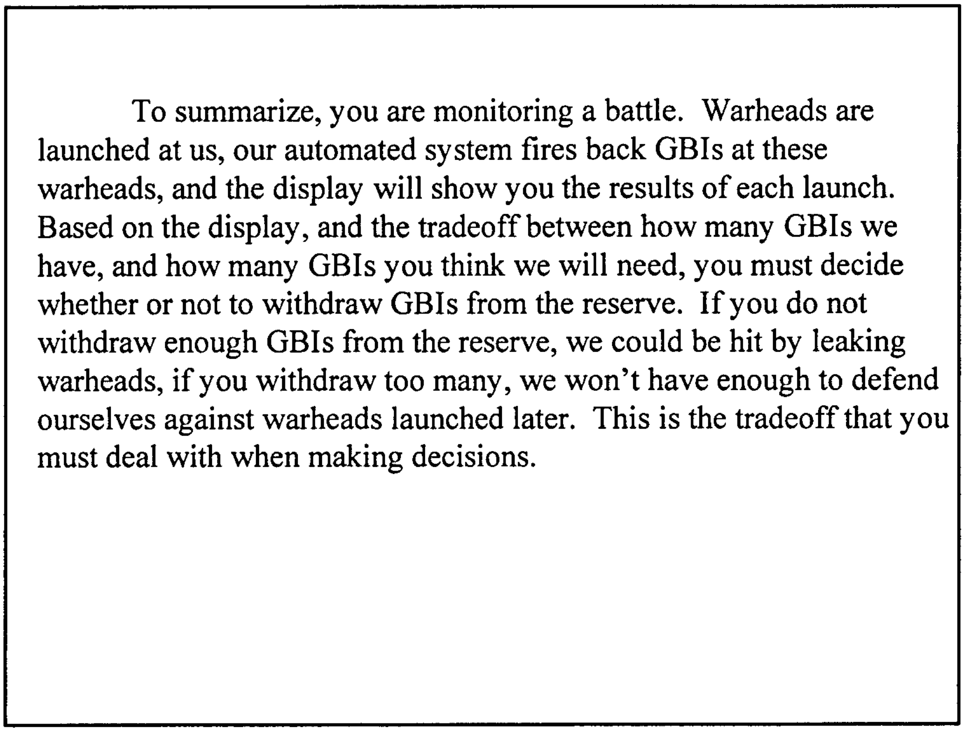


The solid red line and number next to it indicates the current number of GBIs on hand at any given time. Notice the green area between the red line and the line corresponding to the worst case. The green area means that even if the worst case occurs, you will still have enough GBIs to handle the threat.

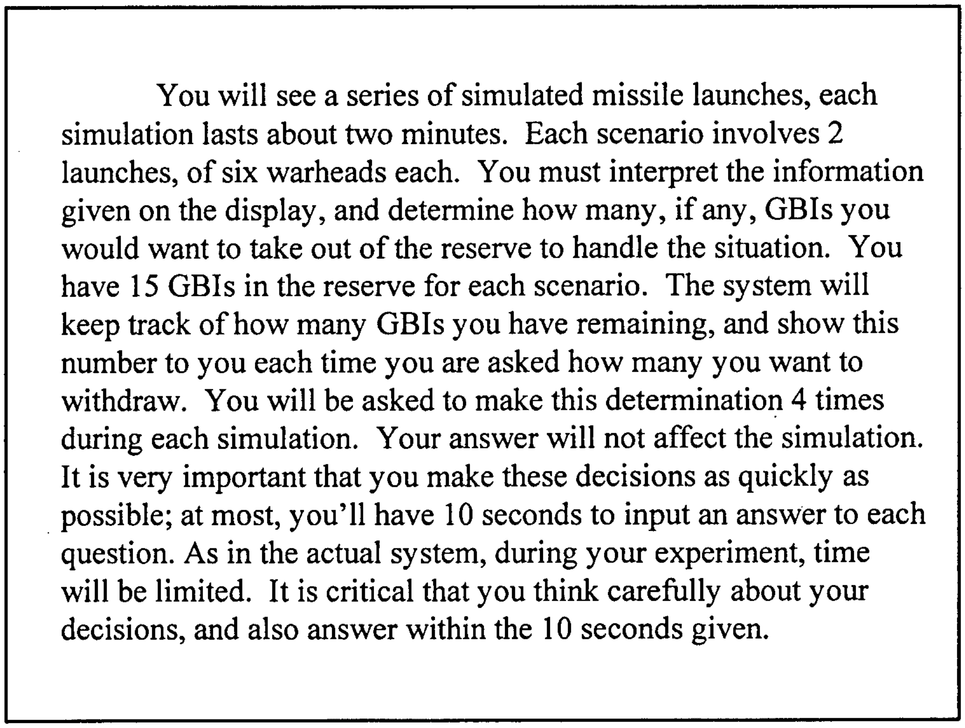


As the attack evolves, you must continue to make assessments of how much at risk we are in failing to destroy all incoming warheads, given how many GBIs we have. (Remember, the supply of GBIs starts at 28 and decreases for every GBI that is launched). The Risk Display provides you with the information necessary to make these risk assessments. If you feel that the risk is too high, or may soon become too high, then you should choose to withdraw GBIs from the reserve when you are given the opportunities to do so. However, you should be careful in deciding to withdraw GBIs since once they are launched, they can not be brought back, and will not be available for possible future attacks. Remember that the number of reserve GBIs is limited to 15.

To reiterate, the NMD system automatically fires back at incoming warheads, this is something you monitor, not control. You do control the launch of the 15 reserve GBIs. These reserve GBIs are not included in your risk display.



To summarize, you are monitoring a battle. Warheads are launched at us, our automated system fires back GBIs at these warheads, and the display will show you the results of each launch. Based on the display, and the tradeoff between how many GBIs we have, and how many GBIs you think we will need, you must decide whether or not to withdraw GBIs from the reserve. If you do not withdraw enough GBIs from the reserve, we could be hit by leaking warheads, if you withdraw too many, we won't have enough to defend ourselves against warheads launched later. This is the tradeoff that you must deal with when making decisions.



You will see a series of simulated missile launches, each simulation lasts about two minutes. Each scenario involves 2 launches, of six warheads each. You must interpret the information given on the display, and determine how many, if any, GBIs you would want to take out of the reserve to handle the situation. You have 15 GBIs in the reserve for each scenario. The system will keep track of how many GBIs you have remaining, and show this number to you each time you are asked how many you want to withdraw. You will be asked to make this determination 4 times during each simulation. Your answer will not affect the simulation. It is very important that you make these decisions as quickly as possible; at most, you'll have 10 seconds to input an answer to each question. As in the actual system, during your experiment, time will be limited. It is critical that you think carefully about your decisions, and also answer within the 10 seconds given.

Finally, it is very important that you think about your decisions carefully. In the actual system, operator decisions would result in saving the lives of American citizens, or the failure to do so.

Each simulation will automatically load following the end of the previous simulation, but they will not start automatically. You must press the ▷ (play) button at the bottom right of your screen when you are ready to start the simulation. You can not stop a simulation once it is in progress. At the end of each scenario, the experimenter will ask you how confident you felt about the correctness of your responses, from 1-10 with 10 being the most confident. Throughout the scenarios, different parts of the display will be highlighted, and different probabilities will be shown to aid your decision making.

Now, we'll walk through a practice scenario to make sure everything is clear. Please ask any questions you might have. Thank you for your participation.

APPENDIX B: POST-EXPERIMENT QUESTIONNAIRE

Subject Number _______

Thank you for completing the experiment, please take a few minutes to answer some questions concerning the design and ease of use of the NMD system. Please circle your response

1. **Overall, how easy to interpret was the NMD computer display?**
   (1=very easy, 7=very difficult)
   1 2 3 4 5 6 7

2. **How understandable were the instructions given before your experiment?**
   (1=not understandable, 7=very understandable)
   1 2 3 4 5 6 7

3. **How easy to interpret was the Risk-Display part of the system?**
   (1=very easy, 7=very difficult)
   1 2 3 4 5 6 7

4. **What piece of information helped you to make your decisions the most?**
   a) best case probability
   b) expected case probability
   c) worst case probability
   d) past results of launches
   e) none of the above

5. **During the experiment, I found myself confused:**
   almost never sometimes often

6. **How comfortable were you in using the NMD computer display to make decisions?**
   (1= not comfortable, 7=very comfortable)
   1 2 3 4 5 6 7

Please list any additional comments about your preferences for the NMD display system or the experiment on the back of this form.